FC and Ethernet Bridging for Converged Networks

Liran Liss
Mellanox Technologies

Agenda

- Why bridging?
- Stateless bridging
- Bridging protocols
- BridgeX
- FCoXX implementation
- EoIB implementation
- SW stack

The Converged-network Promise

- Replace
  - Multiple adaptors
  - Multiple networks
  - Multiple cables
- With
  - Single CNA per host
  - Single wire, single network
- But – we still need to communicate with the outside world…

Bridges to the Rescue

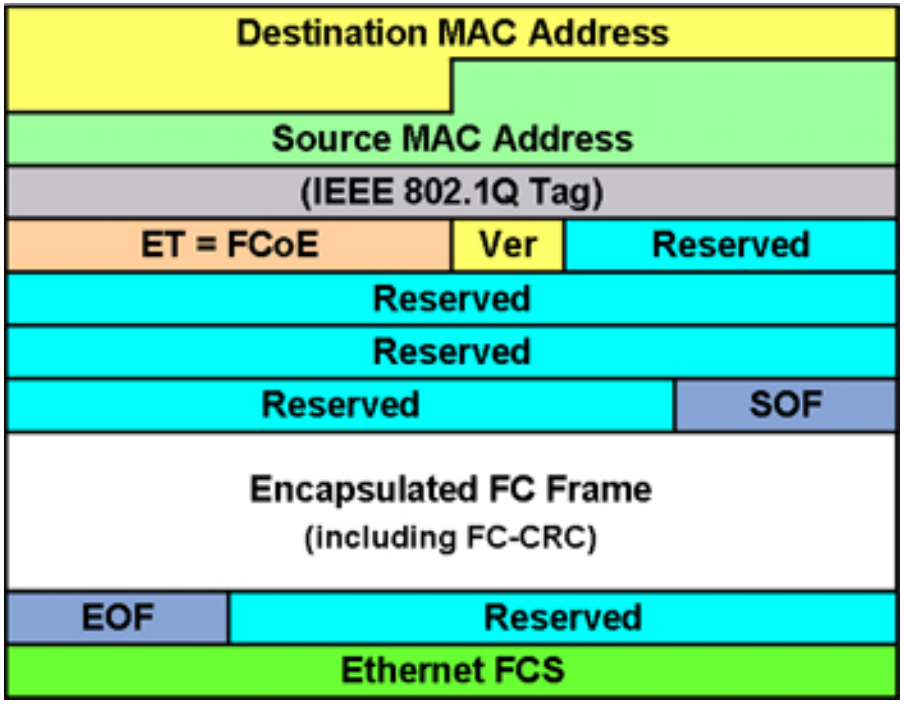
- For CEE networks
  - Ethernet to Fibre Channel bridges
    - Connect servers within the converged fabric to FC SANs
- For InfiniBand networks
  - Infiniband to Fibre Channel bridges
    - Same as above
  - Infiniband to Ethernet bridges
    - Connect servers within the converged fabric to the Internet

Stateless Bridging

- Bridge does not hold any state beyond a single packet
- Servers “speak” the target protocol
- Converged network is only a conduit between the host and the bridge
  - Native protocol PDUs are carried (encapsulated) over the converged fabric protocol
- Advantages:
  - Unlimited scalability (both in load and in number of sessions)
  - Bridges can be simple and cheap
    - Only perform encap/decap operations
- Example: Fibre Channel over Ethernet (FCoE)

FCoE at a Glance

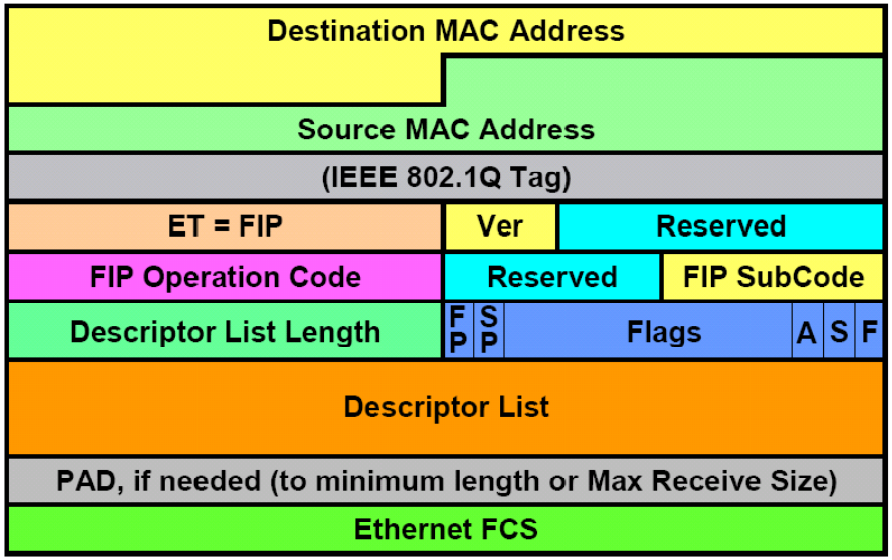
- Mapping of Fibre Channel frames natively over CEE
  - Replaces FC0+FC1 layers of FC stack with Ethernet
  - FCF (the FCoE switch / bridge) does encap/decap between FC←→FCoE
- Seamless integration with FC networks and management SW



FCoE – Discovery

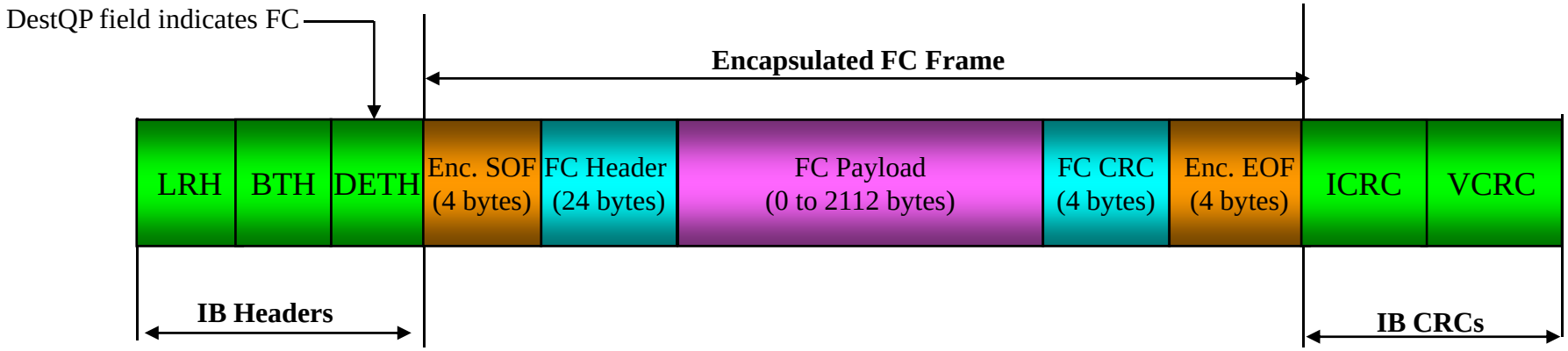
- Achieved by FCoE Initialization Protocol (FIP)
- FIP packets contain
  - Opcode/subcode
  - A descriptor list of TLVs (e.g. FCF-MAC, FC-ID, switch name)
- FIP opcodes:
  - Solicitation / Advertisement
  - Login / login-ack
  - Keep alive / clear virtual link

FIP frame format



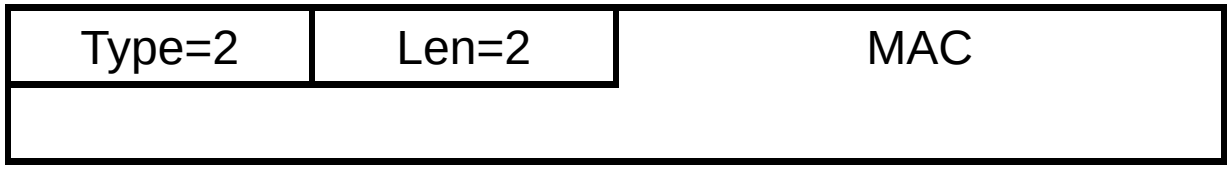
FCoIB

- Modeled after FCoE
  - Replaces FC0+FC1 by InfiniBand UD
    - FC-2 and above remain the same
  - Bridge does encap/decap of IB headers
- Addressing
  - Destination QP indicates FC Nx_Port
  - Gateway maintains mapping between FCID and IB address



FCoIB Discovery

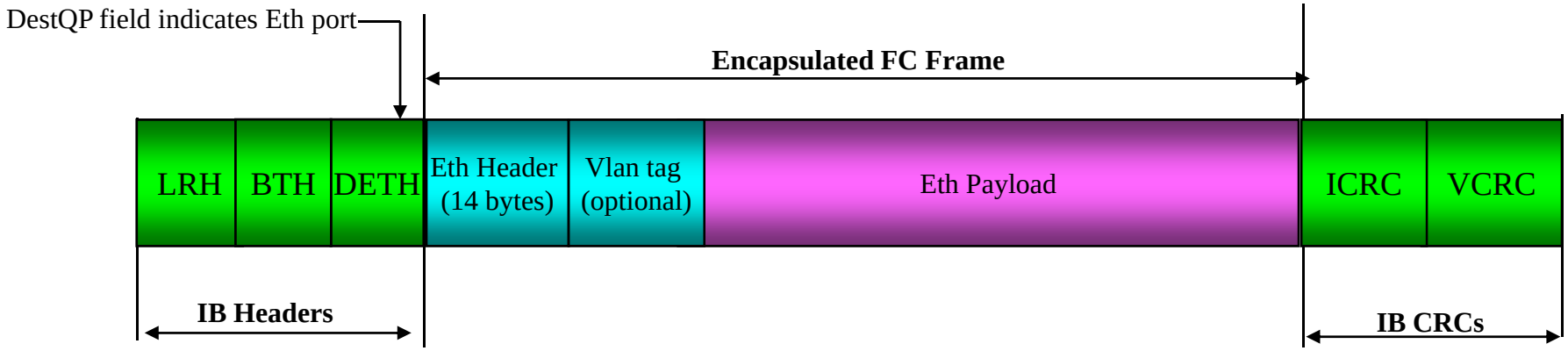
- You guessed it – FIP
- Replace MAC TLV:



- with InfiniBand Address TLV, which contains:
  - LID
  - QPN
  - SL
  - Port GUID

EoIB

- Encapsulates Ethernet frames in InfiniBand UD packets
  - Purely L2 – IP is not involved
  - Bridge does encap/decap
- Addressing
  - Destination QP indicates Eth port for IB→Eth traffic
  - GW maintains mappings between MAC+VLAN and LID+QP for Eth→IB traffic



EoIB Discovery

- FIP-like protocol
  - Solicitation/advertisement
  - Login / Login-ack
- Additional FIP opcode for transferring context tables
  - More on this later

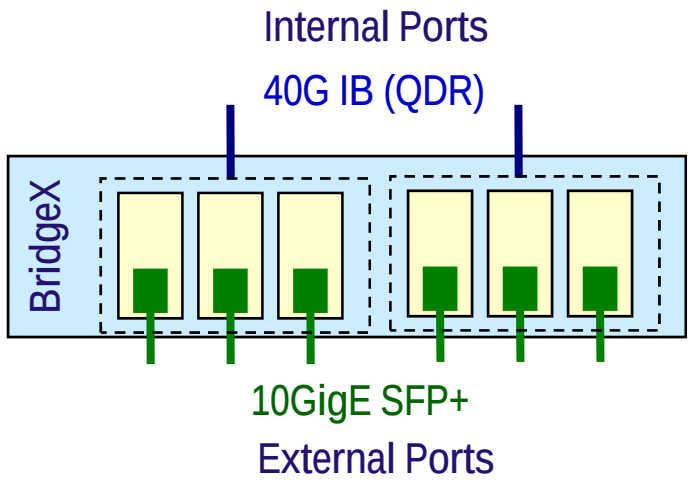
BridgeX

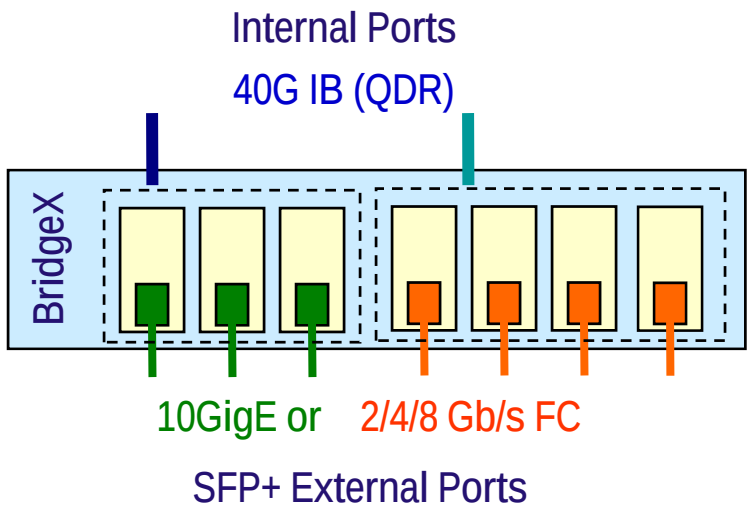
- Multi-protocol bridge
  - FCoE
  - FCoIB
  - EoIB
- Stateless
  - Simple encap/decap operation
  - Highly scalable
    - “Limited” only by wire-speed (for any packet size)
- Virtual Protocol Interconnect (VPI) support
  - Internal network: SDR/DDR/QDR IB, XAUI/SFP+ 10GE
  - External network: 10GE SFP+, 2/4/8G FC

MTB4020 top of rack bridge system

IB to FC/Ethernet Configurations

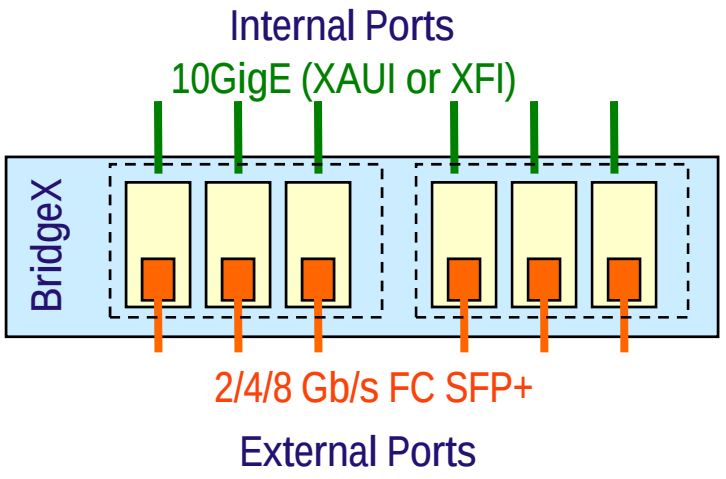
- Internal ports: SDR/DDR/QDR IB
- External port options:
  - 6 x 10GigE
  - 8 x FC
  - 4 x FC + 3x 10GigE

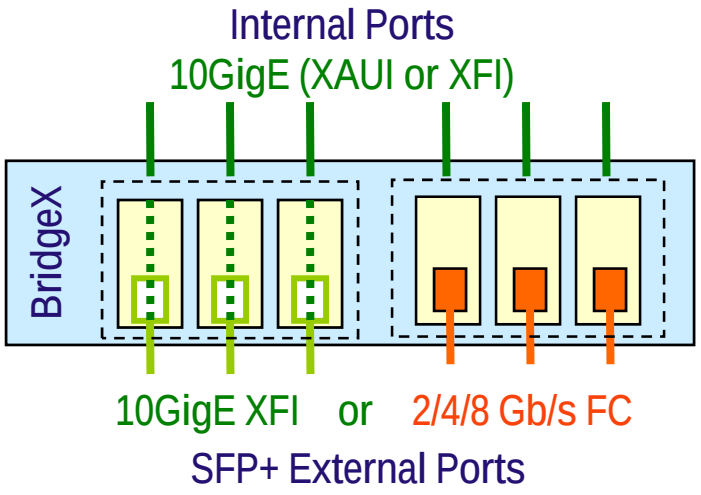




Ethernet to FC Configurations

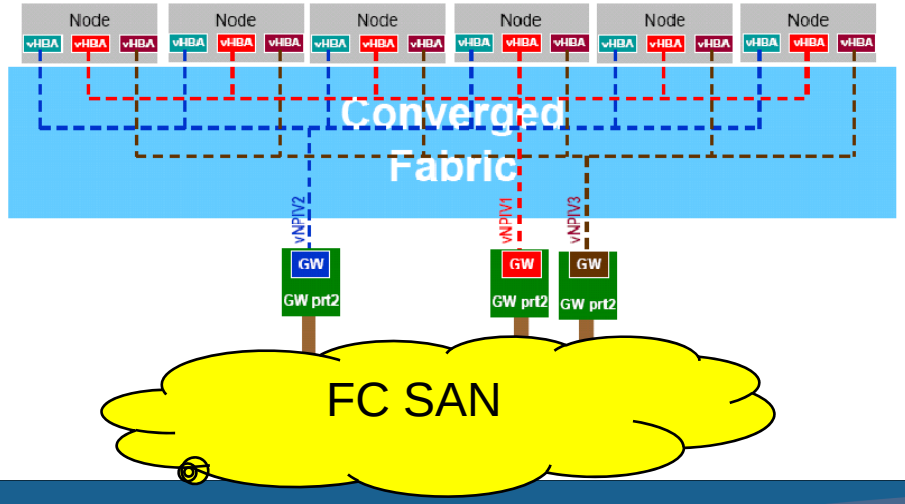
- Internal ports: 10GE
- External port options:
  - 6 x FC
  - 6 x 10GigE
  - 3 x FC + 3 x 10GigE





FCoXX Implementation

- Each external port is associated with a Virtual NPIV (vNPIV) set
  - Represents FC connectivity within the internal network
  - Appears as a normal NPIV port to SAN
  - Holds a context table containing a record for each vHBA
- GW logs in to SAN and translates host FLOGI packets to FDISC

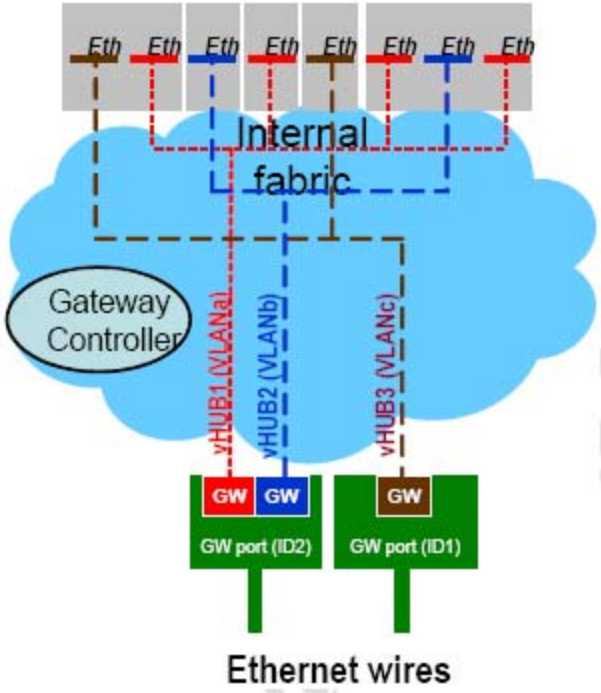


FCoXX – Data Path Operation

- Host
  - Egress path – destination address is always the gateway implementing vNPIV end-point
  - Ingress path – FC Frame is received as a payload and presented to the vHBA on the host
- Gateway
  - Ingress from FC
    - Gateway looks up the D_ID and sends an encapsulated frame to the host
    - If lookup fails, packet is dropped
  - Egress from vHBA
    - Data frames: sent out on the FC port after source validation
    - Control frames: gateway executes the login

EoIB - implementation

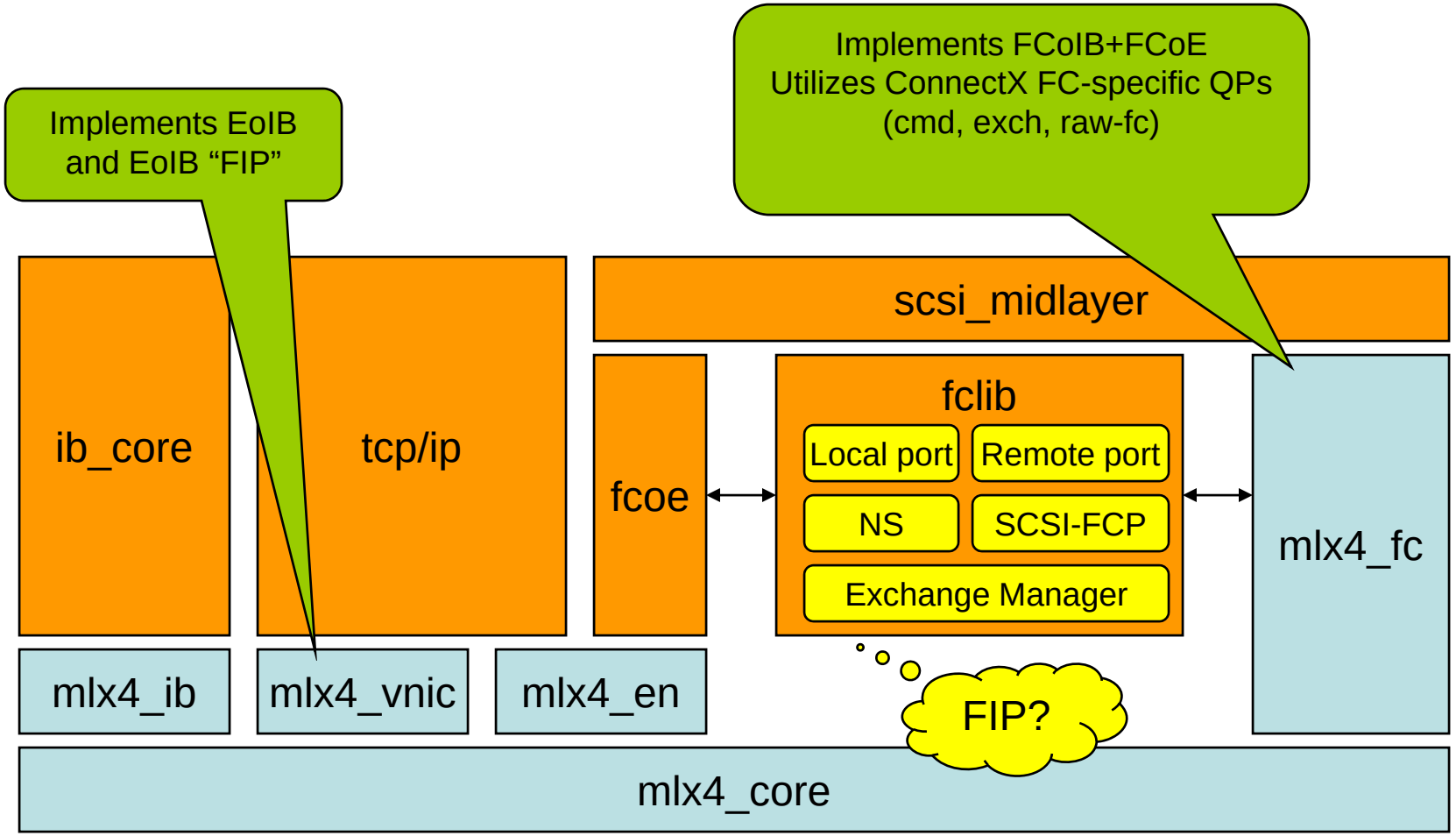
- Each external port is associated with one or more Virtual Hubs (vHubs)
- A vHub
  - Represents an Ethernet broadcast domain within the internal network (VLAN)
  - Holds a context table containing a record for each vNIC
    - The context table is distributed to all vNICs
  - Associated with 3 IB multicast groups
    - Broadcast
    - Context table updates
    - Context table distribution



EoIB – Data-path Operation

- Host
  - EoIB driver looks up the destination address {DMAC, VLAN} on the vHub context table
    - If the destination matches than the InfiniBand address of the matching entry is used
    - If it misses, then the gateway address is used
  - Received frames are delivered to the Ethernet netdevice
- Gateway
  - Egress packets received by the gateway from a vNIC are delivered to the corresponding Ethernet port
  - Ingress packets are associated to a vHub based on VLAN ID, and then sent to the corresponding host

Software Stack



Summary

- Bridges are required to connect the converged fabric to the outside world
- FCoIB and EoIB follow FCoE concepts
  - Discovery based on FIP
- Using a stateless bridging approach, BridgeX achieves this with
  - High performance
  - High scalability
  - Low cost