OPENFABRICS ALLIANCE

Ethernet Services over IPoIB

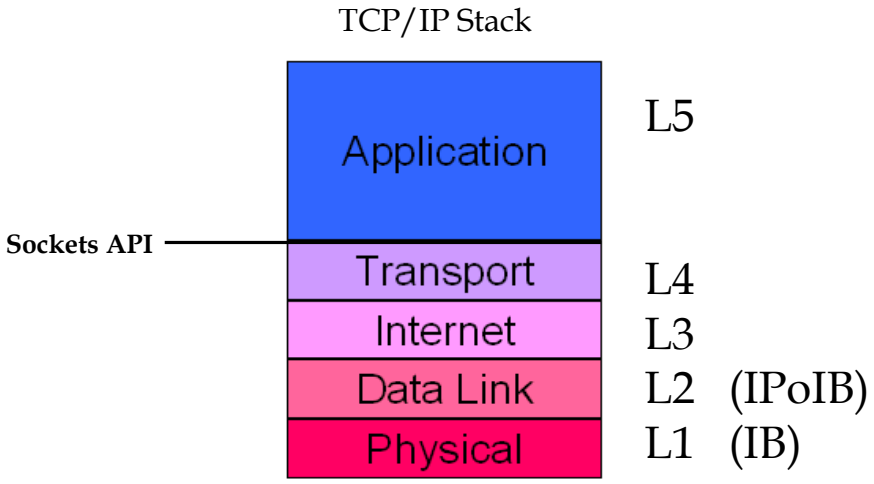
Ali Ayoub, Mellanox Technologies
March, 2012

Background

- What is IPoIB?
  - IP encapsulation over InfiniBand
  - RFC 4391/4392
  - Provides IP services over InfiniBand fabric
- Benefits:
  - Acts like an data-link within the TCP/IP Stack
    - Socket-based apps run transparently
  - Allow users to run IP-based unmodified applications on InfiniBand
  - Supports NIC offloads: CSUM, TSO, etc.. And other performance features such as: NAPI, TSS, RSS, etc..

IPoIB Packet Format

- IPoIB integrated as Data Link layer
- Sockets API is not affected



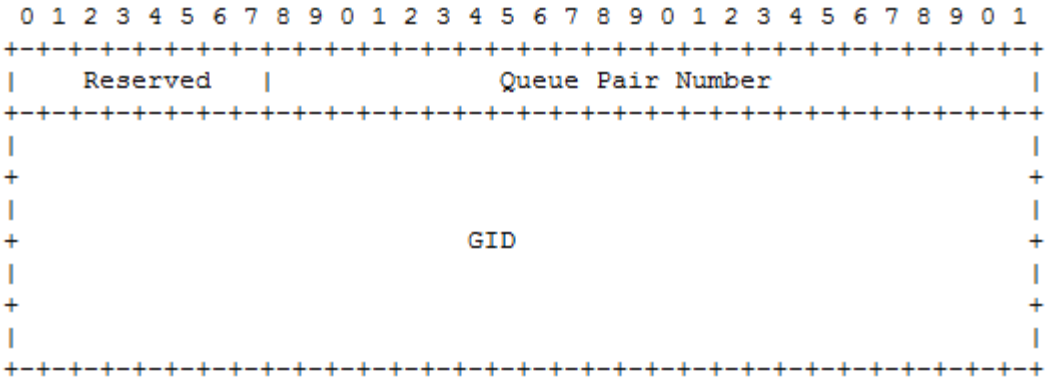
- Packet frame:

Limitations

- IPoIB is limited for IP applications only
  - No Ethernet Header encapsulation
- IPoIB network interface doesn't act like a standard Ethernet Network Device
    - Non-standard Ethernet interface utilities & characteristics
    - 20 Bytes link-layer address
    - Ethernet link-layer (MAC) address is 6 Bytes
    - DHCP requires using dhcp-client identifier
  - Host administrator must be aware of PKEYs
    - While Ethernet interfaces support VLANs
    - vconfig command cannot be used

Limitations cont.

- Link Layer setting, modification, migration
  - IPoIB Link Layer address is based on QPN/GID
    - Cannot be controlled by the user
  - Host administrator cannot set or re-configure the link layer address (MAC)

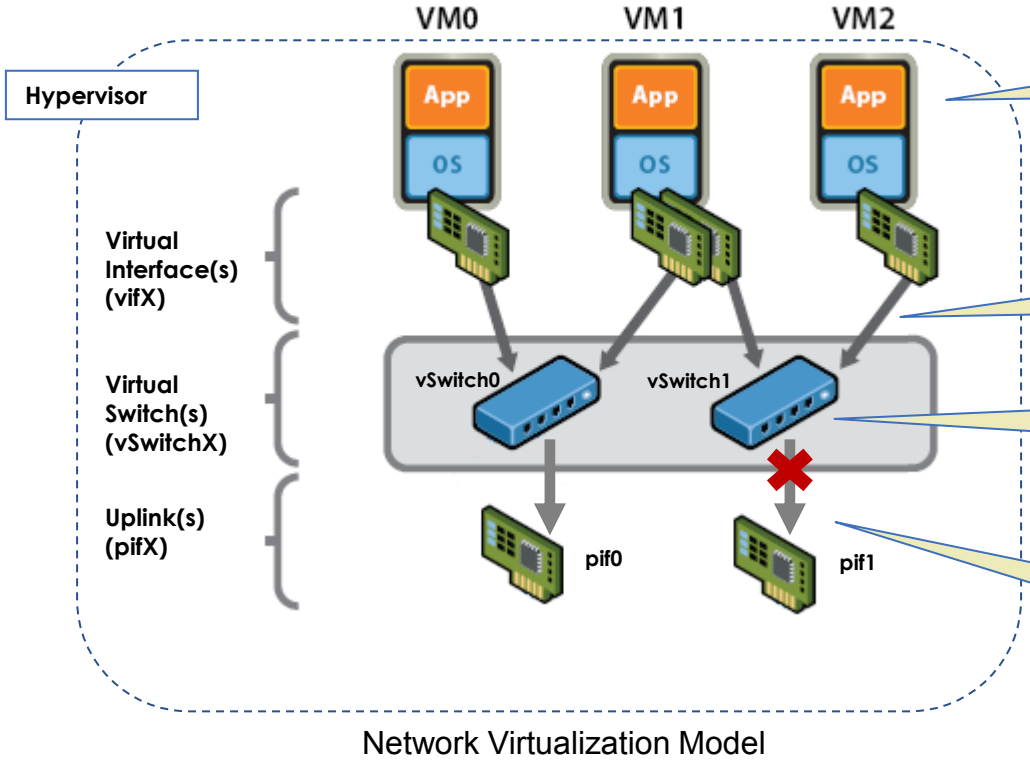


IPoIB Link Layer (RFC 4391)

Limitations cont.

- IPoIB cannot be used in fully-virtualized or para-virtualized environments
  - Hypervisor networking model normally use “bridged mode”
  - Virtual Switch (vSwitch) is Ethernet L2 switch
  - IPoIB NIC cannot be enslaved to a vSwitch
- Promiscuous mode is unsupported
  - Simplifies vSwitch functionality

IPoIB & Network Virtualization



Network Virtualization Model

VM "sees" an Ethernet Network Device (e1000, virtio, vmxnet, etc..)

Packet sent out by the VM includes an Ethernet header

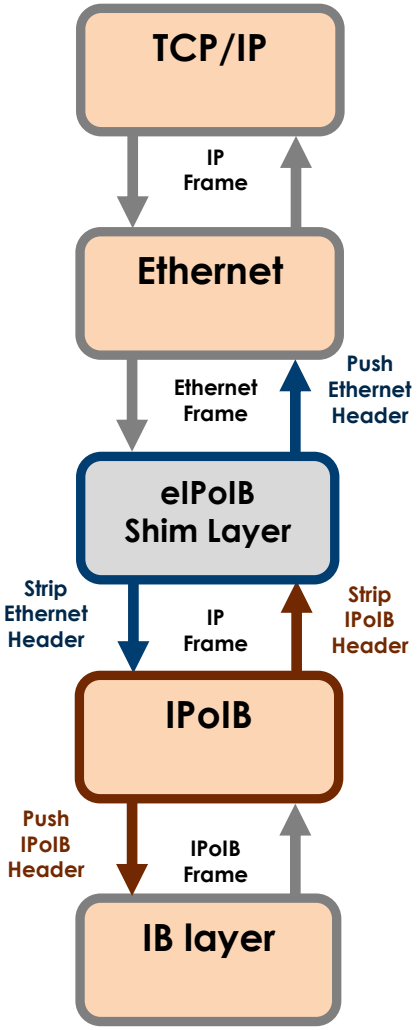
vSwitch forwards the Ethernet packet to the host Network Interface

Native IPoIB driver serves only IP packets (doesn't expect Ethernet Header)

Reference http://www.vmware.com/resources/techresources/997

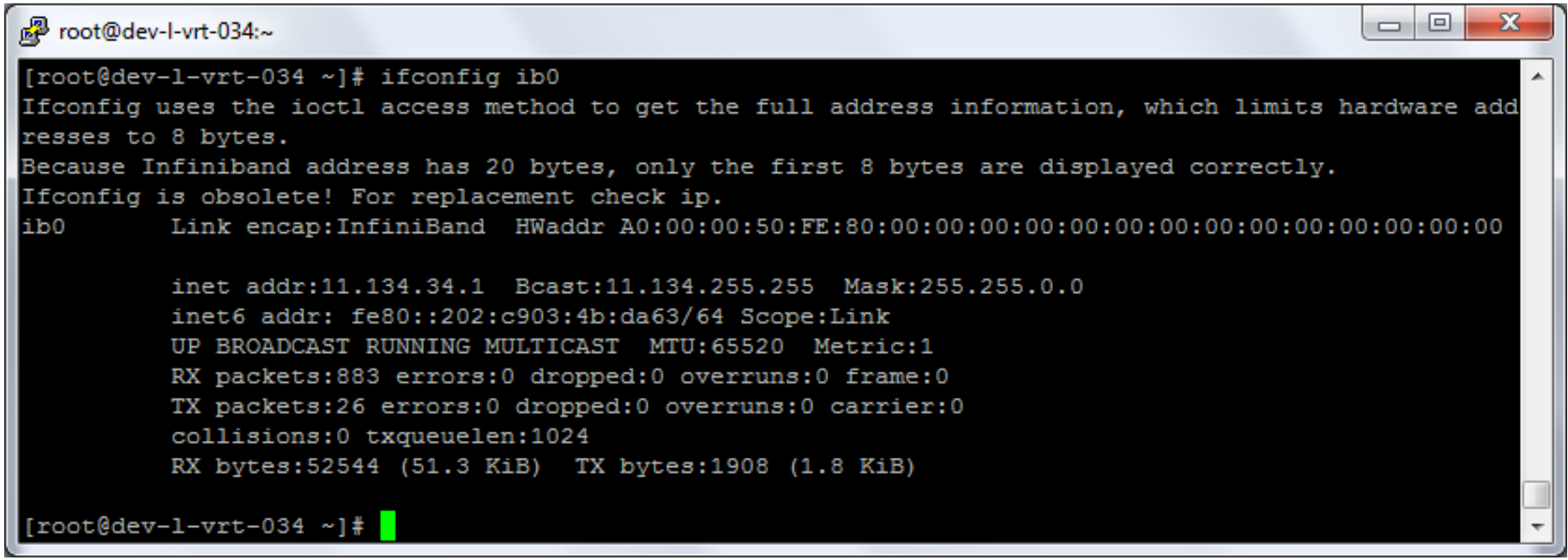
Solution

- Create a new Ethernet Network Device over IPoIB interface
- Managed by “eIPoIB” kernel module
- Registers a standard Ethernet Interface into the Operating System
- Same “Look & Feel” as an Ethernet NIC
  - ifconfig, vconfig, ethtool, promiscuous, etc..
- eIPoIB is a **shim layer** driver between the TCP/IP stack and the native IPoIB



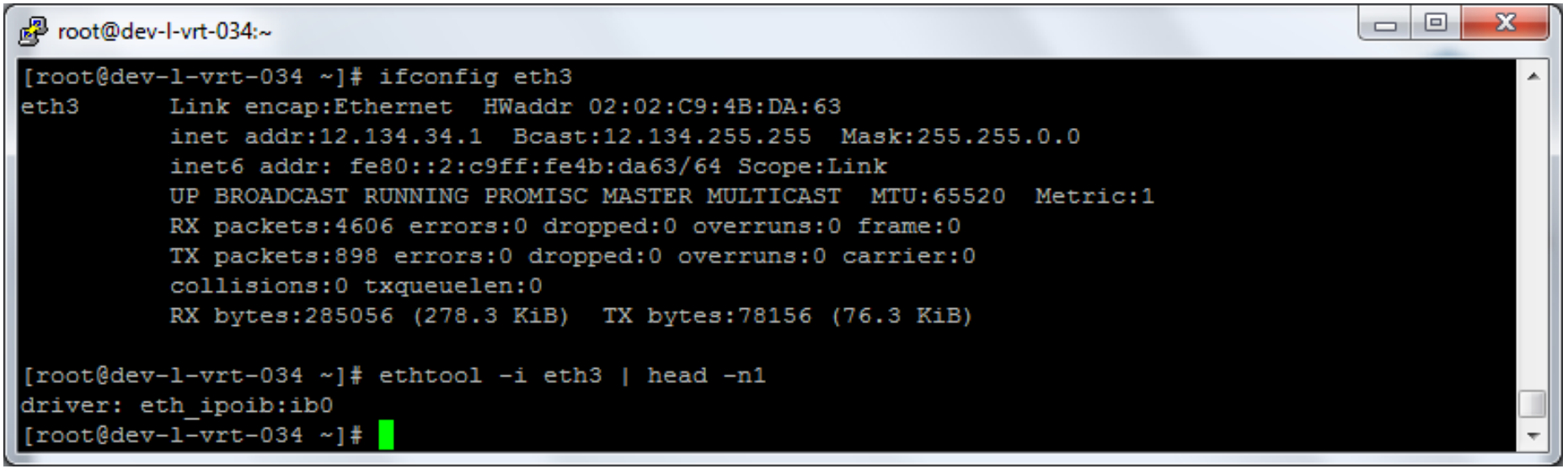
eIPoIB vs. IPoIB Interface

- ifconfig ib0



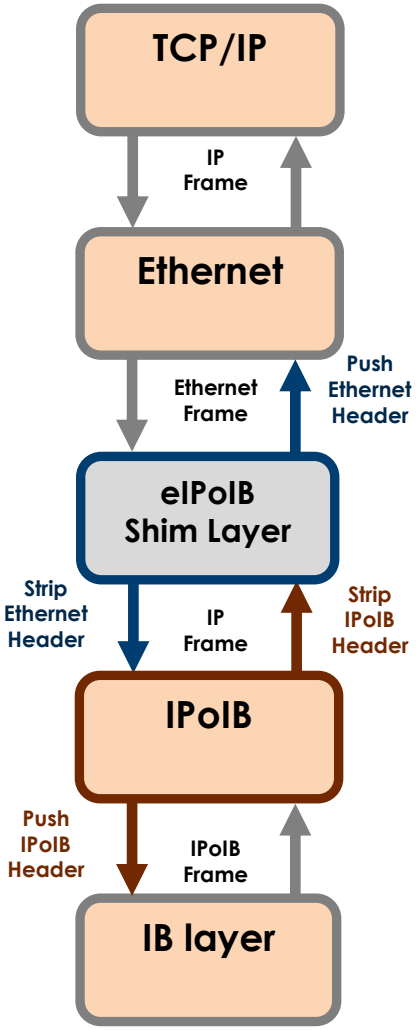
eIPoIB vs. IPoIB Interface

- ifconfig eth3

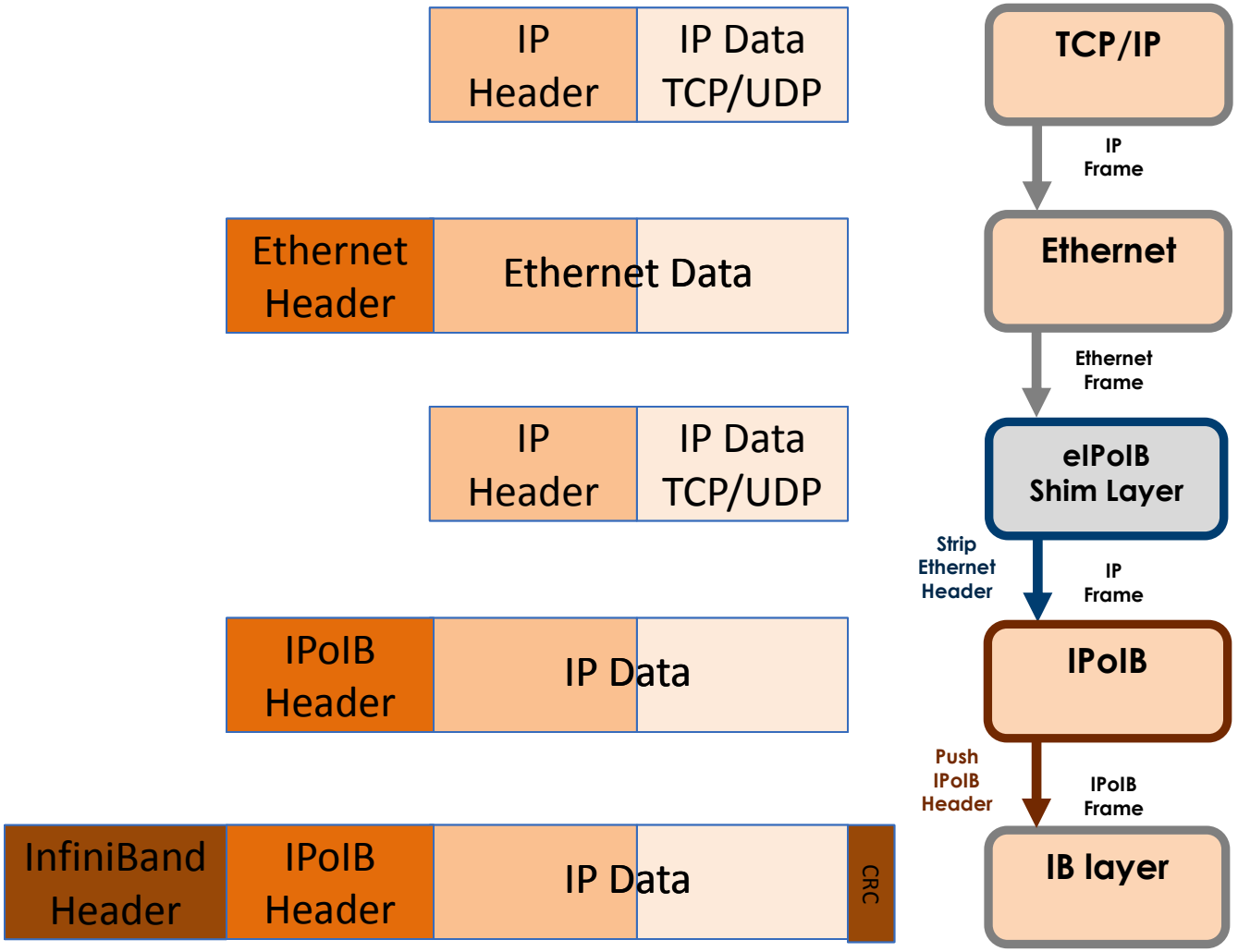


eIPoIB Operations in a Nutshell

- Initialization:
  - Register Ethernet network interface (eth0)
  - Map it to native IPoIB interface (ib0)
- Transmit:
  - Packet to be sent has Ethernet Header
  - eIPoIB strips the Ethernet Header
  - Pass the IP packet to the underlying IPoIB interface.
    - Non IP/ARP/RARP packets are dropped
  - Native IPoIB interface sends the packet out
- Receive:
  - Packet received is an IP packet
  - Native IPoIB handovers the IP packet to eIPoIB layer
  - eIPoIB pushes the Ethernet header
  - Packet forwarded to upper layers as regular Ethernet frame

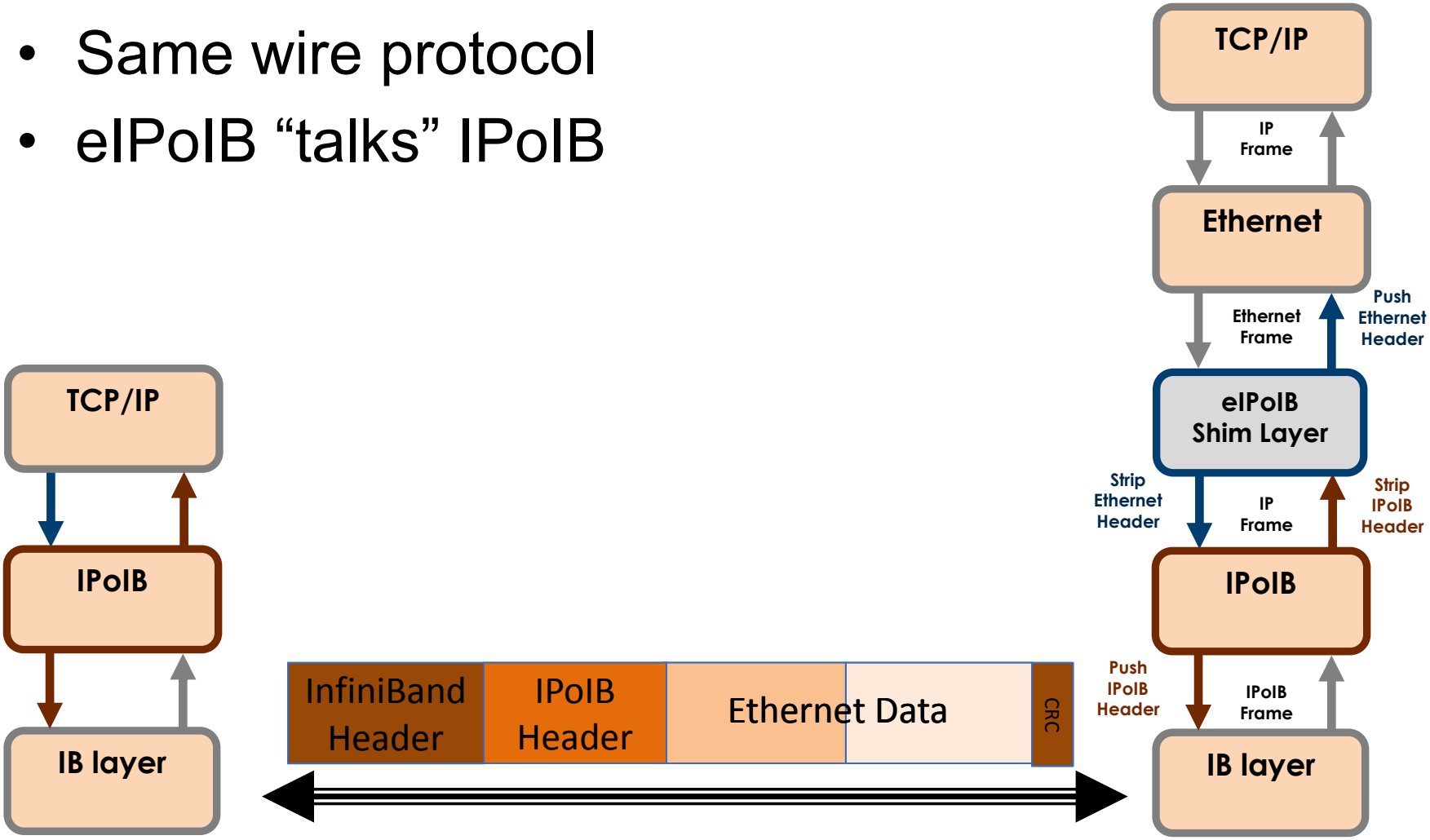


eIPoIB Frame Flow

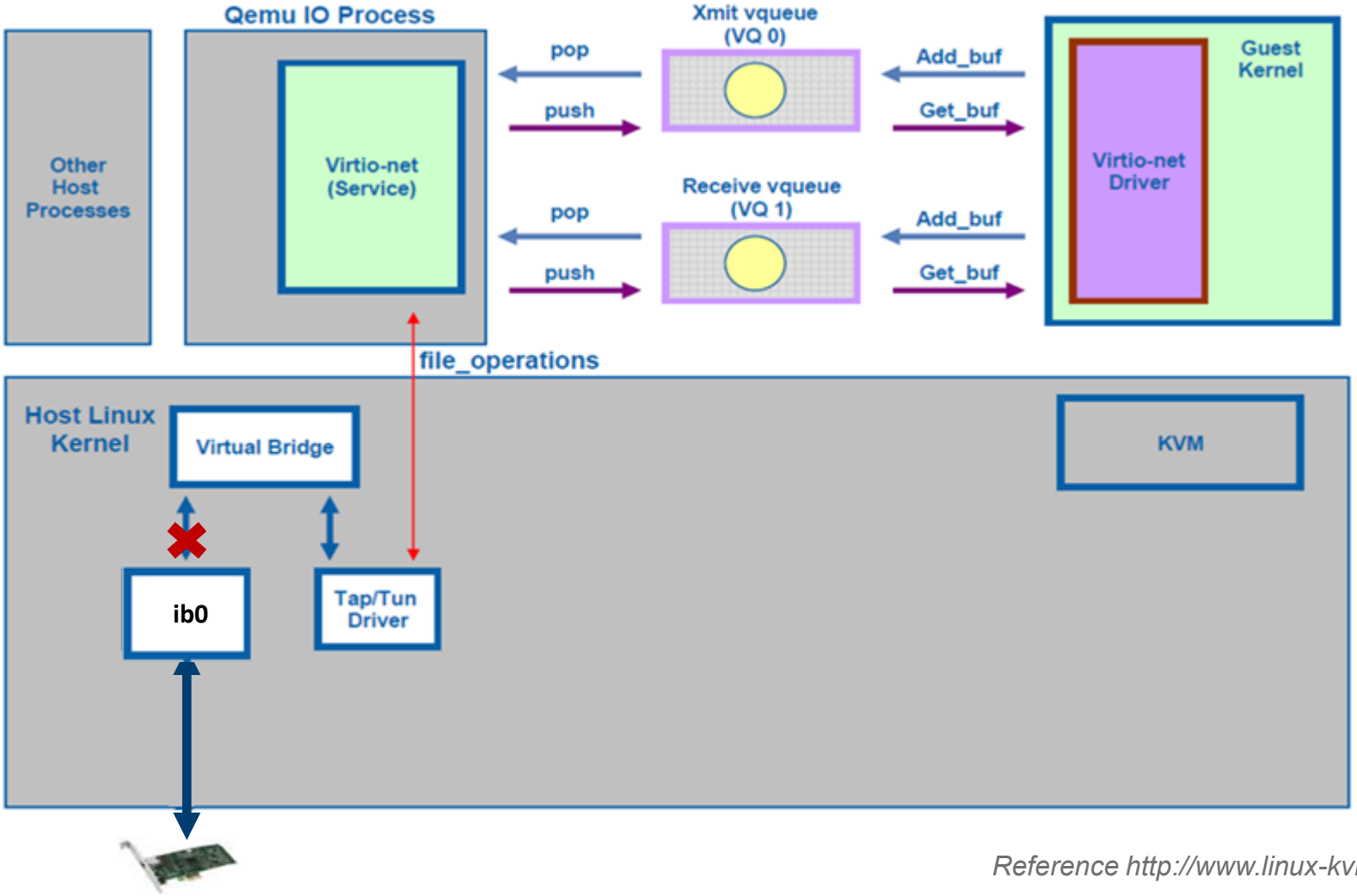


eIPoIB/IPoIB Interoperability

- Same wire protocol
- eIPoIB “talks” IPoIB

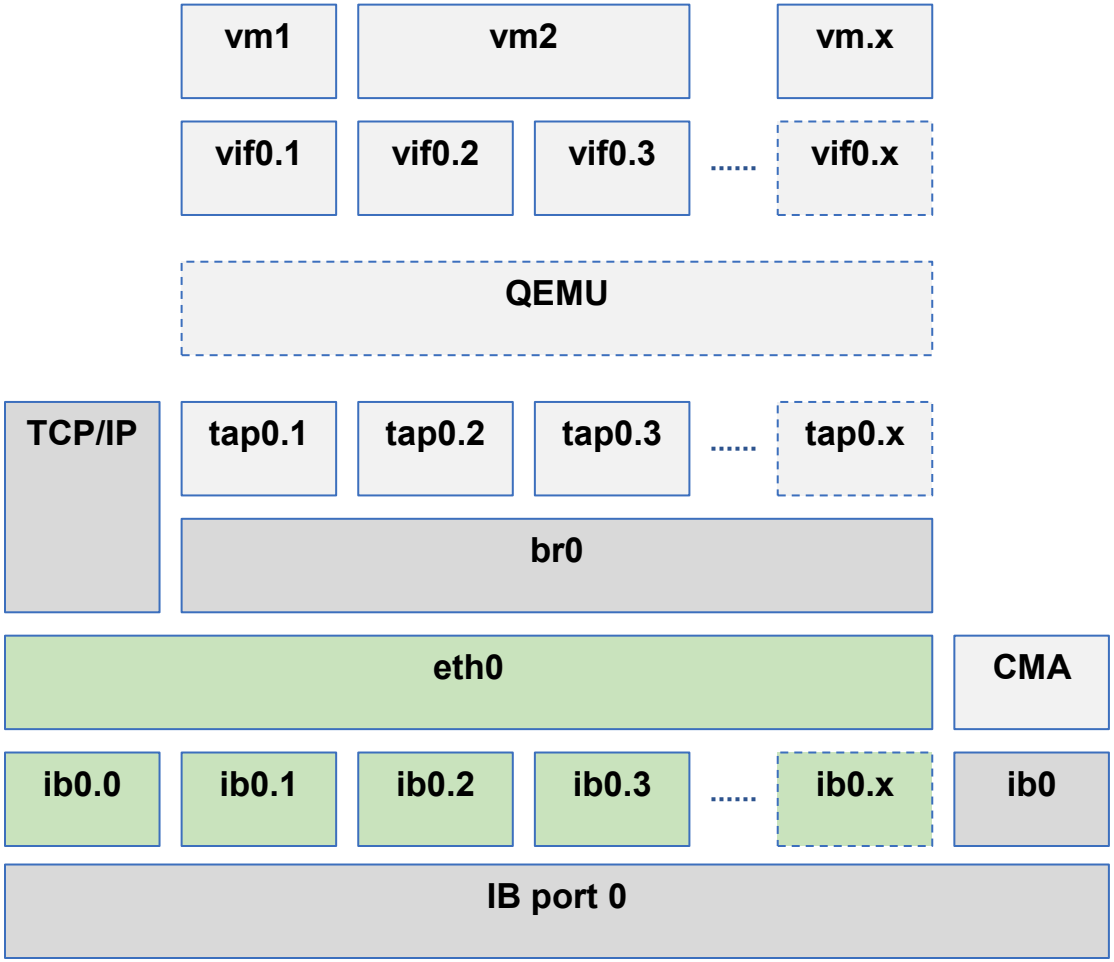


eIPoIB Networking Model (KVM)



Reference http://www.linux-kvm.org/wiki

eIPoIB Design (KVM)



eIPoIB – Closer Look

- Promiscuous mode
- MAC Translation

Promiscuous Mode

- Normally vSwitch uplink is put in promiscuous mode
  - InfiniBand doesn't support promiscuous mode
- Promiscuous mode is simulated by:
  a. Snooping the src.mac and vlan of outgoing packets
  b. OS notifies the driver when a new MAC/VLAN need to be "served". For example: the driver can get a notification from libvirt library (available for KVM/XEN) when a new VM virtual NIC is created
- Multicast promiscuous support requires ***IGMP/MLD Snooping*** in eIPoIB level

MAC Translation

- Requirements:
  - eIPoIB exposes an Ethernet MAC
    - Local Ethernet MAC (LEMAC) 6 bytes length
  - eIPoIB neighbors are seen as Ethernet neighbors
    - Remote Ethernet MAC (REMAC) 6 bytes length
  - IPoIB Local MAC used on the wire
    - Local IPoIB MAC (LIMAC) 20 bytes
  - IPoIB neighbors' MAC used on the wire
    - Remote IPoIB MAC (RIMAC) 20 bytes

MAC Translation cont.

OPENFABRICS ALLIANCE

Receive Flow

- Receiver QP => dst.mac
- SQPN/SLID => src.mac
  - Remember QPN/LID to IPoIB-MAC (QPN/GID) mapping
- IPoIB header => Ethernet ethertype
- Replace IPoIB header by Ethernet header

Transmit Flow

- src.mac => QP (child interface)
  - Source MAC normally controlled by the host/hypervisor admin
- dst.mac => IPoIB-MAC
- Ethernet header ethertype => IPoIB Header
- Strip Ethernet header, and handover packet to IPoIB

ARP/NDP packets are modified in TX/RX flow, so the MAC addresses in the packet payload are updated accordingly

OPENFABRICS ALLIANCE

Questions?