Ethernet over InfiniBand

Ali Ayoub, Mellanox Technologies
April 2013
802.1Q VLAN
WoL/NCSI/vCDNI
IPv4/IPv6

In other words; an eth0 interface that acts like an eth0 interface
Goal

- Seamless Support for Ethernet Services over InfiniBand Network
  - IP and non-IP Applications
  - Virtualization (vSwitch)
  - 802.1Q
- Seamless Ethernet Management
  - DHCP, PXE, etc.
  - Load Balancing & High Availability
    - Unmodified Bonding/Teaming driver support
- Protocol may be distributed
  - Doesn’t rely on central software/hardware manager
- Simple bridging between EoIB and Ethernet
# What’s New

<table>
<thead>
<tr>
<th></th>
<th>Ethernet</th>
<th>EoIB</th>
<th>IPoIB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ethernet Header</td>
<td>Present</td>
<td>Present</td>
<td>Not Present</td>
</tr>
<tr>
<td>Compatibility with L2-based apps</td>
<td>Seamless</td>
<td>Seamless</td>
<td>Not Supported&lt;br&gt;Needs special handling when using eIPoIB</td>
</tr>
<tr>
<td>MAC Setting</td>
<td>Any</td>
<td>Any</td>
<td>Limited: based on QPN and GID</td>
</tr>
<tr>
<td>MAC Length</td>
<td>6 bytes</td>
<td>6 bytes</td>
<td>20 bytes</td>
</tr>
<tr>
<td>Migration</td>
<td>Transparent to the netdev driver</td>
<td>Transparent to the netdev driver</td>
<td>Requires special handling</td>
</tr>
<tr>
<td>VLAN ID</td>
<td>Any</td>
<td>Any</td>
<td>IPoIB: Not Supported&lt;br&gt;eIPoIB: Mapped to PKEY (1..128 only, cannot exceed PKEY range)</td>
</tr>
</tbody>
</table>
Model

• **Ethernet Overlay Network on top of InfiniBand Underlying Network (UD Transport)**
• InfiniBand Network as a “giant” Virtual Ethernet Switch (VES)
• End points may have one or more Virtual Ports (vPort) connected to the VES
• A Virtual NIC (vNIC) represents the Ethernet Interface within the end-point, connected directly to the vPort
• A Gateway (GW) can be implemented the same way as a host with multiple pNIC/vNIC instances
Model

- VES is distributed; each vPort holds a Forwarding Database (FDB) table.
- Optionally, a VES manager can be used to push the FDB table to the end points.
- A Gateway (GW) can be implemented the same as a host with multiple pNIC/vNIC instances.
Packet Format

- **OS**
- **vPort**
- **HW**

**UD Transport** is used for Encapsulation

- InfiniBand Header
- EoIB Encap Header
- Ethernet Header
- Ethernet Data
- CRC

- Ethertype, src/dst MAC, [VLAN]
- version, src/dst vPort,
Address Resolution

• What’s New:
  – Ethernet Link Layer (MAC) is decoupled from the underlying InfiniBand network
    • Allows using any MAC address; a must for virtualization models where the hypervisor is responsible for VM’s MAC setting
  – EoIB is decoupled from ARP/NDP protocols
    • No dependency on the OS address resolution and Control Plane
    • Allows EoIB to have its own Control Plane and carry information/notifications not available in ARP/NDP
  – Learning
Address Resolution

• How it works:
  – Each end-point holds a Forwarding Database (FDB) table
  – The FDB is used to map the Ethernet packet based on MAC/VLAN to the corresponding InfiniBand Address Handle
  – FDB is updated based on ingress traffic learning as well as EoIB Control Plane
  – If mapping is missing, the packet is flooded (distributed mode)
    • Similar to VXLAN approach

• Egress Packet Flow:
FDB

• Construction:
  – Learn incoming traffic to map MAC/VLAN to a vPort
    • Same approach as physical Switch learning
  – Use EoIB Control Plane (vPort Request/Reply) to map vPort to IB Address
  – SA query is sent out to get the PathRecord based on the IB Address

• Scheme

<table>
<thead>
<tr>
<th>Overlay Address</th>
<th>Underlying Address</th>
<th>Physical Address</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAC VLAN vPort ID</td>
<td>QPN LID [GID]</td>
<td></td>
</tr>
<tr>
<td>MAC VLAN vPort ID</td>
<td>QPN LID [GID]</td>
<td></td>
</tr>
<tr>
<td>MAC VLAN vPort ID</td>
<td>QPN LID [GID]</td>
<td></td>
</tr>
<tr>
<td>MAC VLAN vPort ID</td>
<td>QPN LID [GID]</td>
<td></td>
</tr>
</tbody>
</table>
Ping Example

Learn MAC/VLAN to vPort mapping

Learn vPort to IB address mapping

SA Query is issued* to obtain the PathRecord

Learn MAC/VLAN to vPort mapping

FDB is incomplete; flooding is used (in distributed mode)

Learn vPort to IB address mapping

SA Query is issued* to obtain the PathRecord

* Not shown in the diagram
Thank You
Backup
Layers

Inner Layer
Encapsulation Header
Outer Layer

Overlay Packet
Encapsulation Header
Underlay Layer
Physical Layer

Ethernet
Encapsulation Header

Overlay key: mac/vlan/tni
Underlay key: vPortID
Physical Key: QPN/LID/GID
SA Query
Multicast

Table 19: Multicast GiD Layout

| 31 | 30 | 29 | 28 | 27 | 26 | 25 | 24 | 23 | 22 | 21 | 20 | 19 | 18 | 17 | 16 | 15 | 14 | 13 | 12 | 11 | 10 | 9  | 8  | 7  | 6  | 5  | 4  | 3  | 2  | 1  | 0  |
|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|
|    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
|    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
|    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
|    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
|    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |

- **Prefix**
  - Offset: 00h
  - PKEY
  - DMAC

- **DMAC**
  - Offset: 04h

- **Version**
  - Offset: 08h
- **Type**
- **NS**
- **Reserved0**
- **VID**
- **CH**
VES Instances

• Each PKEY defines a VES instance
• VES can serve multiple VLANs
  – VLAN and PKEY are decoupled
  – The administrator can limit the use of specific VLAN group for each VES instance for higher security