Support Infiniband Link Layer

Shahaf Shuler
DPDK Summit Userspace - Dublin- 2017
Agenda

- Why Infiniband in DPDK?
- Infiniband Intro
- Infiniband network addressing
- IPoIB
- POC results
- Upstream Infiniband to DPDK
Why Infiniband?

- Many large scale Infiniband clusters in the HPC market
- Fast packet processing is required also there
  - Parallel Distributed storage applications
- Infiniband was defined with user space direct access from the start
  - DPDK is better optimized for high packet rate than the original verbs API
    - Max rate with verbs: ~30Mpps
    - Max rate with DPDK: ~50Mpps
Infiniband (IB) computer network standard

Centralized subnet management using the SM

Benefits

- Flatter topology
- Low latency – 0.7 usec
- L4 queues
- RDMA and remote atomic operation

RDMA_WRITE(src_va, length, dst_node, dst_va)
Infiniband (IB) computer network standard.

Centralized subnet management using the SM

Benefits

- Flatter topology
- Low latency – 0.6 usec
- L4 queues
- RDMA and remote atomic operation

RDMA_READ(src_va, dst_node, dst_va, length)
# Infiniband Network Addressing

<table>
<thead>
<tr>
<th>Ethernet</th>
<th>Infiniband</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAC (6 Bytes)</td>
<td>LID (2 Bytes)</td>
<td>LIDs are configured by the SM. LIDs are statically configured in the switch.</td>
</tr>
<tr>
<td>VLAN (4 Bytes)</td>
<td>PKey</td>
<td>No VLAN offloads</td>
</tr>
<tr>
<td>IPv4 (4 Bytes) IPv6 (16 Bytes)</td>
<td>GID (16 Bytes)</td>
<td>GIDs are fixed. cannot be changed.</td>
</tr>
<tr>
<td>UDP/TCP port (2 Bytes)</td>
<td>QP (3 Bytes)</td>
<td>Apart from MC, packet targets single QP. No IB RSS yet. No promiscuous No all multi</td>
</tr>
<tr>
<td>ARP</td>
<td>Path query</td>
<td>{LID,MTU} = PQ(GID)</td>
</tr>
</tbody>
</table>

---

**Infiniband Header**

- **Destination Local Identifier (D_ID)**
  - Bits: 32-24, 23-16, 15-8, 7-0
  - Values: 0-3, VL, LVer, SL, Rs/2, LNH

- **Source Local Identifier (S_ID)**
  - Bits: 31-24, 23-16, 15-8, 7-0
  - Values: 0-7, VL, LVer, SL, Rs/2, LNH

- **Packet Length (11 bits)**
  - Bits: 4-7
  - Values: Reserve 5

- **Flow Label**
  - Bits: 4-7
  - Values: PayLen, NextHdr, HopLimit

- **Destination QP**
  - Bits: 31-24, 23-16, 15-8, 7-0
  - Values: 0-7, Destination QP

- **OpCode**
  - Bits: 31-24, 23-16, 15-8, 7-0
  - Values: 0-7, OpCode

- **Partition Key**
  - Bits: 31-24, 23-16, 15-8, 7-0
  - Values: 0-7, Partition Key

- **PSN - Packet Sequence Number**
  - Bits: 8-11
  - Values: Reserved 7, Reserved 8 (masked in CRC)
- IPoIB – encapsulation of IP packet in IB message.
- Linux generic netdev for InfiniBand
- It is possible to do RSS and promiscuous (all IPs)
- It is possible to use TSO and checksum offloads

```
3  ip address show
2: ib0: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 4092 qdisc mq state UP qlen 1024
    link/infinitand a0:00:03:00:fe:80:00:00:00:00:00:00:02:09:03:00:02:ff:ff:ff:ff:ff:ff:ff:ff:ff:ff:ff:ff:ff
    inet 50.50.50.12/24 brd 50.50.50.255 scope global ib0
        valid lft forever preferred lft forever
    inet6 fe80::202:e903:2f:ff:ff:fffe:64 scope link
        valid lft forever preferred lft forever
3: eth3: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 1500 qdisc mq state UP qlen 1000
    link/ether e61d:2dfe:feed:34ac:64 brd ff:ff:ff:ff:ff:ff:ff:ff
    inet 40.40.40.12/24 brd 40.40.40.255 scope global ens3
        valid lft forever preferred lft forever
    inet6 fe80::e61d:2dfe:feed:34ac:64 scope link
        valid lft forever preferred lft forever
```
POC Results

- **POC**
  - Done with Weka.io (Parallel distributed Storage App based on DPDK)
  - DPDK v17.05 patched with IPoIB support
  - Replacing SM query with udp socket IB address exchange inside the application with the PMD help
  - rte_flow rules based on well-known udp port to steer traffic to the PMD queues

- **HW**
  - ConnectX-4, single port, speed 56Gb/sec
  - Intel(R) Xeon(R) CPU E5-2643 v4 @ 3.40GHz. Cluster size – 20 nodes.

<table>
<thead>
<tr>
<th></th>
<th>Single core</th>
<th>2 cores</th>
</tr>
</thead>
<tbody>
<tr>
<td>Uni-dir</td>
<td>45.3 Gbps</td>
<td>50Gbps</td>
</tr>
<tr>
<td>(262x4KB from one to other)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bidir</td>
<td>28.36Gbps</td>
<td>39.2Gbps (per direction)</td>
</tr>
<tr>
<td>(262x4KB both ways)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Discussion: Upstream Infiniband to DPDK

- `rte_ib_dev`?
- Sub section inside `ethdev` for `IPoIB`?
- Helper libraries for IB address resolution?
- `Mbuf` fields change
  - Application sets headers starting from `IPoIB` header
  - PMD needs to add the IB headers
Questions?

Shahaf Shuler
shahafs@mellanox.com
GUID = IPoIB_ARP (IP)

Kernel service

{LID, MTU} = PATH_RECORD(GUID)

Infiniband CM has Socket semantic API (listen, connect)

User-space application listen and respond with the QP parameters