Intel Omni-Path (OPA) Virtual Network Interface Controller (VNIC)

Intel Omni-Path (OPA) Virtual Network Interface Controller (VNIC) feature supports Ethernet functionality over Omni-Path fabric by encapsulating the Ethernet packets between HFI nodes.

Architecture

The patterns of exchanges of Omni-Path encapsulated Ethernet packets involves one or more virtual Ethernet switches overlaid on the Omni-Path fabric topology. A subset of HFI nodes on the Omni-Path fabric are permitted to exchange encapsulated Ethernet packets across a particular virtual Ethernet switch. The virtual Ethernet switches are logical abstractions achieved by configuring the HFI nodes on the fabric for header generation and processing. In the simplest configuration all HFI nodes across the fabric exchange encapsulated Ethernet packets over a single virtual Ethernet switch. A virtual Ethernet switch, is effectively an independent Ethernet network. The configuration is performed by an Ethernet Manager (EM) which is part of the trusted Fabric Manager (FM) application. HFI nodes can have multiple VNICs each connected to a different virtual Ethernet switch. The below diagram presents a case of two virtual Ethernet switches with two HFI nodes:

```
                            +-------------------+
                            |      Subnet/      |
                            |     Ethernet      |
                            |      Manager      |
                            +-------------------+
                               /          /
                             /           /
                           /            /
                         /             /
+-----------------------------+  +------------------------------+
|  Virtual Ethernet Switch    |  |  Virtual Ethernet Switch     |
|  +---------+    +---------+ |  | +---------+    +---------+   |
|  | VPORT   |    |  VPORT  | |  | |  VPORT  |    |  VPORT  |   |
+--+---------+----+---------+-+  +-+---------+----+---------+---+
         |                \        /                 |
         |                  \    /                   |
         |                    \/                     |
         |                   /  \                    |
         |                 /      \                  |
     +-----------+------------+  +-----------+------------+
     |   VNIC    |    VNIC    |  |    VNIC   |    VNIC    |
     +-----------+------------+  +-----------+------------+
     |          HFI           |  |          HFI           |
     +------------------------+  +------------------------+
```

The Omni-Path encapsulated Ethernet packet format is as described below.

Bits	Field
Quad Word 0:	
0-19	SLID (lower 20 bits)
20-30	Length (in Quad Words)
31	BECN bit
32-51	DLID (lower 20 bits)
52-56	SC (Service Class)
57-59	RC (Routing Control)
60	FECN bit
61-62	L2 (=10, 16B format)
63	LT (=1, Link Transfer Head Flit)
Quad Word 1:	
0-7	L4 type (=0x78 ETHERNET)
8-11	SLID[23:20]
12-15	DLID[23:20]
16-31	PKEY
32-47	Entropy
48-63	Reserved
Quad Word 2:	
0-15	Reserved
16-31	L4 header
32-63	Ethernet Packet
Quad Words 3 to N-1:	

Bits	Field
0-63	Ethernet packet (pad extended)
Quad Word N (last):	
0-23	Ethernet packet (pad extended)
24-55	ICRC
56-61	Tail
62-63	LT (=01, Link Transfer Tail Flit)

Ethernet packet is padded on the transmit side to ensure that the VNIC OPA packet is quad word aligned. The 'Tail' field contains the number of bytes padded. On the receive side the 'Tail' field is read and the padding is removed (along with ICRC, Tail and OPA header) before passing packet up the network stack.

The L4 header field contains the virtual Ethernet switch id the VNIC port belongs to. On the receive side, this field is used to de-multiplex the received VNIC packets to different VNIC ports.

Driver Design

Intel OPA VNIC software design is presented in the below diagram. OPA VNIC functionality has a HW dependent component and a HW independent component.

The support has been added for IB device to allocate and free the RDMA netdev devices. The RDMA netdev supports interfacing with the network stack thus creating standard network interfaces. OPA_VNIC is an RDMA netdev device type.

The HW dependent VNIC functionality is part of the HFI1 driver. It implements the verbs to allocate and free the OPA_VNIC RDMA netdev. It involves HW resource allocation/management for VNIC functionality. It interfaces with the network stack and implements the required net_device_ops functions. It expects Omni-Path encapsulated Ethernet packets in the transmit path and provides HW access to them. It strips the Omni-Path header from the received packets before passing them up the network stack. It also implements the RDMA netdev control operations.

The OPA VNIC module implements the HW independent VNIC functionality. It consists of two parts. The VNIC Ethernet Management Agent (VEMA) registers itself with IB core as an IB client and interfaces with the IB MAD stack. It exchanges the management information with the Ethernet Manager (EM) and the VNIC netdev. The VNIC netdev part allocates and frees the OPA_VNIC RDMA netdev devices. It overrides the net_device_ops functions set by HW dependent VNIC driver where required to accommodate any control operation. It also handles the encapsulation of Ethernet packets with an Omni-Path header in the transmit path. For each VNIC interface, the information required for encapsulation is configured by the EM via VEMA MAD interface. It also passes any control information to the HW dependent driver by invoking the RDMA netdev control operations:

```
+-------------------+ +----------------------+
|                   | |       Linux          |
|     IB MAD        | |       Network        |
|                   | |        Stack         |
+-------------------+ +----------------------+
         |               |          |
         |               |          |
+----------------------------+      |
|                            |      |
|      OPA VNIC Module       |      |
|  (OPA VNIC RDMA Netdev     |      |
|     & EMA functions)       |      |
|                            |      |
+----------------------------+      |
            |                       |
            |                       |
   +------------------+             |
   |     IB core      |             |
   +------------------+             |
            |                       |
            |                       |
+--------------------------------------------+
|                                            |
|      HFI1 Driver with VNIC support         |
|                                            |
+--------------------------------------------+
```