25GbE Intel® Ethernet
Advanced Features for NFV

Helin Zhang, Intel®
Jingjing Wu, Intel®
Agenda

• Key Hardware Features
• Dynamic Device Personalization (DDP)
• Generic Flow API
• Virtual Function Daemon (VFD)
• Good Performance
• Adaptive Virtual Function (AVF)
Key Hardware Features

- Intel® Ethernet Network Adapter XXV710
  - Single and dual 10/25GbE ports
  - PCIe v3.0, x8
  - New addition to Intel® Ethernet 700 Series

- Network Virtualization Offloads
  - VXLAN, NVGRE, GENEVE, VXLAN-GPE with NSH, MPLS, and more

- Input Set for RSS and Intel® Ethernet Flow Director (FD)
  - Up to 24 of 56 words can be selected

- 3 HASH Algorithms
  - Toeplitz, Simple XOR, Symmetric Simple XOR
## Key Hardware Features for Virtualization

<table>
<thead>
<tr>
<th>Feature</th>
<th>Intel® 82599EB 10 Gigabit Ethernet Controller</th>
<th>Intel® Ethernet 700 Series</th>
</tr>
</thead>
<tbody>
<tr>
<td>SR-IOV support</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>VF to PF mailbox</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Max Number of Virtual functions</td>
<td>64 per port (single queue)</td>
<td>128 per device (globally)</td>
</tr>
<tr>
<td>Max number of Queues</td>
<td>128</td>
<td>1536</td>
</tr>
<tr>
<td>Max number of queues per VF</td>
<td>8</td>
<td>16</td>
</tr>
<tr>
<td>Max number of queues per VMDq2 VSI</td>
<td>8</td>
<td>16</td>
</tr>
<tr>
<td>Max Number of VMDq2 ports</td>
<td>64 per port (single queue)</td>
<td>256 per device (globally)</td>
</tr>
<tr>
<td>MAC addresses</td>
<td>128 per port</td>
<td>1024 per device (globally)</td>
</tr>
<tr>
<td>VLAN tags</td>
<td>64 per port</td>
<td>512 per device (global)</td>
</tr>
<tr>
<td>Queuing to Pool/VSI method</td>
<td>SA or VLAN or (SA and VLAN)</td>
<td>SA, VLAN pairs or SA or VLAN</td>
</tr>
<tr>
<td>Cloud filter in Switch</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>RSS per VF</td>
<td>No (Single RSS used for all VFs).</td>
<td>Yes</td>
</tr>
<tr>
<td>Switching modes</td>
<td>VEB*</td>
<td>VEB, VEPA</td>
</tr>
<tr>
<td>Promiscuous modes per VM</td>
<td>Multicast</td>
<td>VLAN, Multicast, Unicast</td>
</tr>
</tbody>
</table>
Internal Packet Processing

Intel® 82599EB 10 Gigabit Ethernet Controller
- Fixed packets Parse graphic
- Input set of filtering/steering is fixed

Intel® Ethernet 700 Series
- Configurable input set for RSS and Intel Ethernet Flow Director (Intel Ethernet FD)
- Dynamic Device Personalization (DDP) to support more protocol steering
Dynamic Device Personalization (DDP)

- By default, it supported limited protocols, due to hardware resources
  - e.g. VXLAN, GENEVE, NVGRE, MPLS, VXLAN-GPE
- Loadable profiles for packet classification for extra protocols
  - e.g. MPLSoGRE, GTP-U/GTP-C, PPoE
- Configurable tunnel filters for traffic steering
  - Steering packets to VM queues on QinQ/tunnel ID
Generic Flow API Support

- A generic way to configure the hardware
  - Don’t need to know the HW specific filters

- Flow rule
  - Attributes
  - Matching pattern
  - Actions

- Rule management
  - rte_flow_validate()
  - rte_flow_create()
  - rte_flow_destroy()
  - rte_flow_flush()
Example

• Direct the VXLAN packet with specific inner MAC and VNI to queue #2.

```c
const struct rte_flow_item pattern[] = {
  {RTE_FLOW_ITEM_TYPE_ETH, NULL, NULL, NULL},
  {RTE_FLOW_ITEM_TYPE_IPV4, NULL, NULL, NULL},
  {RTE_FLOW_ITEM_TYPE_UDP, NULL, NULL, NULL},
  {RTE_FLOW_ITEM_TYPE_VXLAN, {.vni = 1}, NULL, {.vni = "\xff\xff\xff"}},
  {RTE_FLOW_ITEM_TYPE_ETH,
   .dst = {0x00, 0x11, 0x22, 0x33, 0x44, 0x55}}, NULL,
   .dst = {0xFF, 0xFF, 0xFF, 0xFF, 0xFF, 0xFF}}},
  {RTE_FLOW_ITEM_TYPE_END, NULL, NULL, NULL},
};
const struct rte_flow_action actions[] = {
  {RTE_FLOW_ACTION_TYPE_PF, NULL},
  {RTE_FLOW_ACTION_TYPE_QUEUE, {.index = 2}},
  {RTE_FLOW_ACTION_TYPE_END, NULL},
};
struct rte_flow_error flow_err;
flow_err = rte_flow_create(port_id, NULL, pattern, actions, &flow_err);
```
Virtual Function Daemon (VFD) Support

- DPDK PF driver supports both DPDK and kernel VF
- Lots of VF management features are added
  - Mailbox messages management are added
    - VF requests can be accepted/rejected by VFD
- Kernel PF driver does not support those features
- Enabled on DPDK PF driver for
  - Intel® Ethernet 500 Series (ixgbe)
  - Intel® Ethernet 700 Series (i40e)
- Refer to https://github.com/att/vfd
## Performance

### 2 Cards

- **Traffic Generator I/OA**
- **Intel® Ethernet Network Interface Card X710-DAC2**
- **Socket1**
- **Chipset**
- **PortA**
- **PortB**

### 1 Cards

- **Traffic Generator I/OA**
- **Intel® Ethernet Network Interface Card X710-DAC2**
- **Socket1**
- **Chipset**
- **PortA**
- **PortB**

<table>
<thead>
<tr>
<th>Packet Size (Bytes)</th>
<th>Wire Speed (Mpps)</th>
<th>Packet Rate (Mpps)</th>
<th>%Wire Speed</th>
</tr>
</thead>
<tbody>
<tr>
<td>64</td>
<td>37.2</td>
<td>35.63</td>
<td>95.78%</td>
</tr>
<tr>
<td>128</td>
<td>21.1</td>
<td>21.1</td>
<td>100%</td>
</tr>
<tr>
<td>256</td>
<td>11.3</td>
<td>11.3</td>
<td>100%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Packet Size (Bytes)</th>
<th>Wire Speed (Mpps)</th>
<th>Packet Rate (Mpps)</th>
<th>%Wire Speed</th>
</tr>
</thead>
<tbody>
<tr>
<td>64</td>
<td>37.2</td>
<td>18.1</td>
<td>48.67%</td>
</tr>
<tr>
<td>128</td>
<td>21.1</td>
<td>17.26</td>
<td>81.74%</td>
</tr>
<tr>
<td>256</td>
<td>11.3</td>
<td>10.5</td>
<td>92.98%</td>
</tr>
<tr>
<td>512</td>
<td>5.87</td>
<td>5.71</td>
<td>97.27%</td>
</tr>
<tr>
<td>1024</td>
<td>3.03</td>
<td>2.91</td>
<td>97.32%</td>
</tr>
</tbody>
</table>

Adaptive Virtual Function (AVF)
AVF -- Adaptive Virtual Function

Needs:
- A single VF driver for all generations of Devices.

Solution:
- Adaptive Virtual Function
  - Base features
  - Negotiated Advanced Features

Benefits:
- Existing VM Images will run on the new hardware with no change.

From:
- Intel® Ethernet 700 Series
AVF – HW upgrade

Gen 1

Gen 2

Gen N

Hardware upgrade
AVF -- Adaptive Virtual Function

• Base mode supported
  • Single device ID
  • Support for single level checksum and TSO offload
  • Multi-queue support
  • RSS

• Advanced features
  • Advanced feature introduced by new generation HW.
  • Negotiate with PF driver to expose.
Preserving in Hardware and Software

- **Preserving Base mode**
  - Fixed Minimum Register definition
  - A fixed Meta data format for DMA
  - A Hardware generic mailbox to talk to the PF
  - A Software defined Virtual channel layered on top of Hardware mailbox for expansion

- **Room for expansion**
  - Uncompromising on the base functionality.
  - A large range for hot path registers (Queue and Interrupt)
  - Expandable Virtual channel capability negotiation over the agreed upon communication channel between PF and VF.
  - More advanced features would be added with new drops of AVF driver if the underlying HW device supports.
  - Intel is working on the AVF specification.
Key Takeaways

• 25GbE speed, and better hardware capability
• Generic, flexible and configurable flow classification
• NFV enabled with VFD
• Good performance
• Adaptive VF driver for all Intel® Ethernet 700 Series Network Adapter
End

- Helin Zhang, helin.zhang@intel.com
- Jingjing Wu, jingjing.wu@intel.com