Corporate Update

OpenVswitch hardware offload over DPDK

DPDK summit 2017
Agenda

- ASAP2-Flex for vSwitch/vRouter acceleration
- HW classification offload concept
- OVS-DPDK using HW classification offload
- RFC OVS-DPDK using HW classification offload
- Vxlan in OVS DPDK
- Multi-table
- vxlan HW offload concept
- Rte flow groups - multi-table
ASAP²-Flex for vSwitch/vRouter acceleration

- Offload some elements of the data-path to the NIC, but not the entire data-path
  - Data will still flow via the vSwitch
  - Para-Virtualized VM (not SR-IOV)

- Offloads (examples)
  - Classification offload
    - Application provide flow spec and flow ID
    - Classification done in HW and attach a flow ID in case of match
    - vSwitch classify based on the flow ID rather than full flow spec
    - rte flow is used to configure the classification
  - VxLAN Encap/decap
  - VLAN add/remove
  - QoS
HW classification offload concept

- For every OVS flow DP-if should use the DPDK filter (or TC) to classify with Action tag (report id) or drop.
- When receive use the tag id instead of classify the packet
- for Example:
  - OVS set action Y to flow X
    - Add a flow to tag with id 0x1234 for flow X
    - Config datapath to do action Y for mbuf->fdir.id = 0x1234
  - OVS action drop for flow Z
    - Use DPDK filter to drop and count flow Z
    - Use DPDK filter to get flow statistic
For every datapath rule we add a rte_flow with flow Id.
The flow id cache contain mega flow rules
When packet received with flow id. No need to classify the packet to get the rule
### RFC Performance

- Code submitted by Yuanhan Liu.
- Single core for each pmd, single queue,

<table>
<thead>
<tr>
<th>Case</th>
<th>#flows</th>
<th>Base MPPs</th>
<th>Offload MPPs</th>
<th>improvement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wire to virtio</td>
<td>1</td>
<td>5.8</td>
<td>8.7</td>
<td>50%</td>
</tr>
<tr>
<td>Wire to wire</td>
<td>1</td>
<td>6.9</td>
<td>11.7</td>
<td>70%</td>
</tr>
<tr>
<td>Wire to wire</td>
<td>512</td>
<td>4.2</td>
<td>11.2</td>
<td>267%</td>
</tr>
</tbody>
</table>
Vxlan in OVS DPDK

- There are 2 level of switch that are cascade
- The HW classification accelerate only the lower switch (br-phys1)
- br-phy1 is a kernel interface for vxlan
- The OVS datapath required to classify the inner packet
The action of a rule can be to go to other table. It can be used to chain classification.

### Table 0

- Match A → flow ID 1
- Match B → drop
- Match C → Table 1
- Match D → Table 1
- Default no flow ID

### Table 1

- Match E → flow ID 2
- Match F → flow ID 3
- Default flow ID 4
vxlan HW offload concept

- If the action is to forward to internal interface add HW rule to point to a table named the internal interface.
- If the in port of the rule is internal port (like vxlan) add rule to the table named of the interface with a flow id.
- When a packet is received with a flow id use the rule even if the in port is internal port.
- A packet that tagged with flow id is a packet that came on physical port and classified according to the outer and the inner.
Vxlan HW offload

- If in port is HW port add rule to the HW action can be flow id or to table according to the port to forward to
- If the in port is internal port (like vxlan) add a rule to all the HW port with action flow id. (because traffic can came form any external/HW port)
- The flow id need to be unique.
Rte flow groups - multi-table

- **rte_flow_create()**
  - Group are used in order to add a rule to a table.
  - Need to add new action go to group (RTE_FLOW_ACTION_TYPE_GROUP)

- Table/Group create is implicit
- The user/app need add a default (lowest priority) rule to steer the traffic to a Q and not to continue to next group
Thank You