Corporate Update

OpenVswitch hardware offload over DPDK

DPDK summit 2017

Mellanox TECHNOLOGIES

Connect. Accelerate. Outperform.®

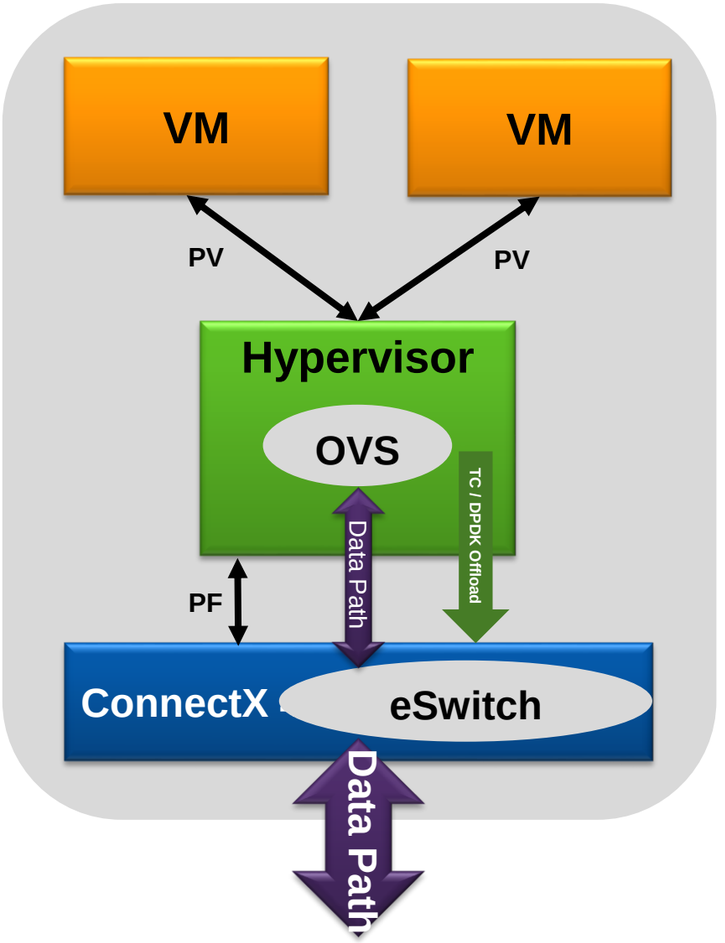
Agenda

- 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

ASAP^2-Flex for vSwitch/vRouter acceleration

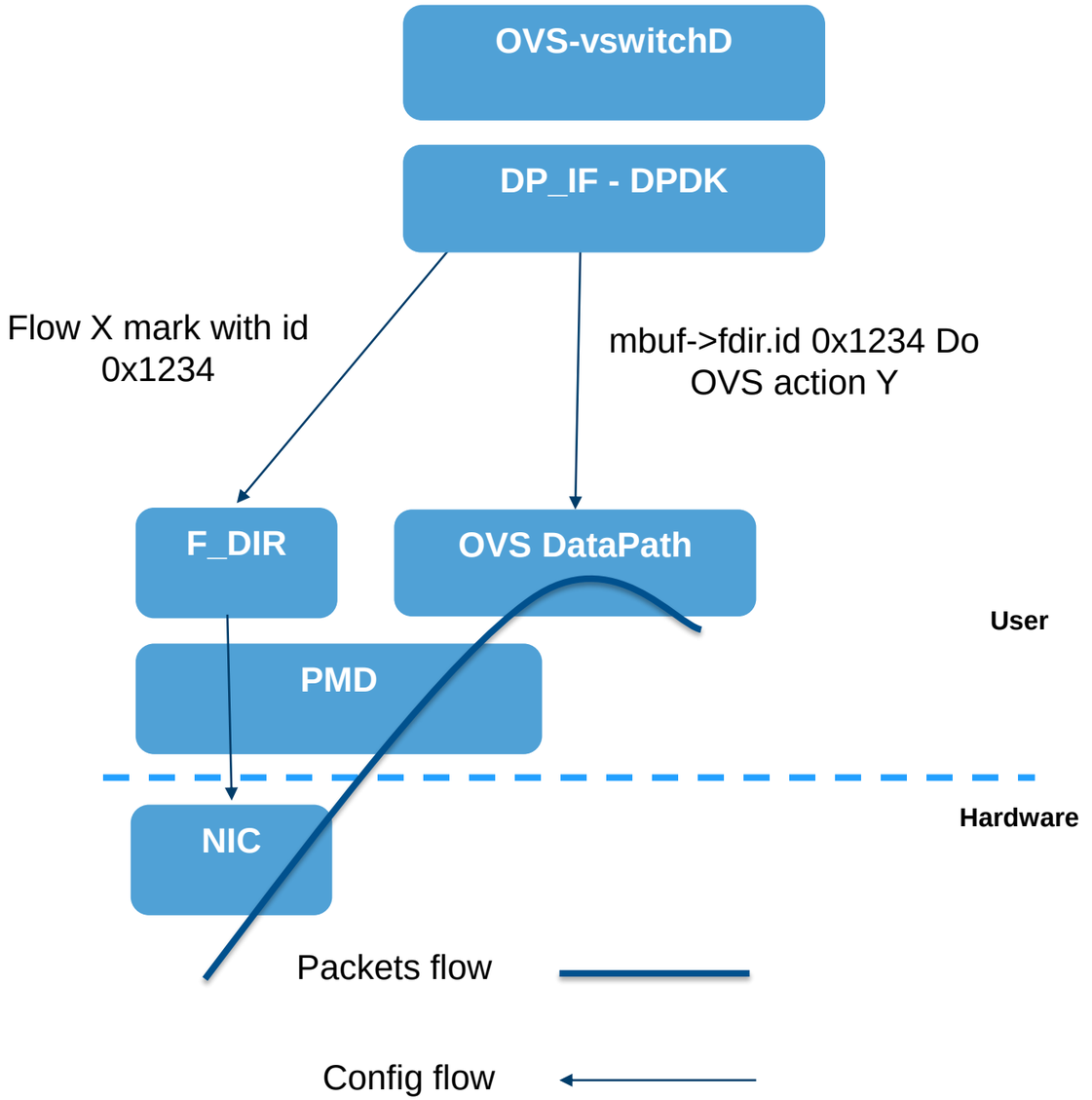
- Offload some elements of the data-path to the NIC, but not the entire data-path
  - Data will still flow via the vSwitch
  - Para-Virtualized VM (not SR-IOV)
- Offloads (examples)
  - Classification offload
    - Application provide flow spec and flow ID
    - Classification done in HW and attach a 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

vSwitch acceleration



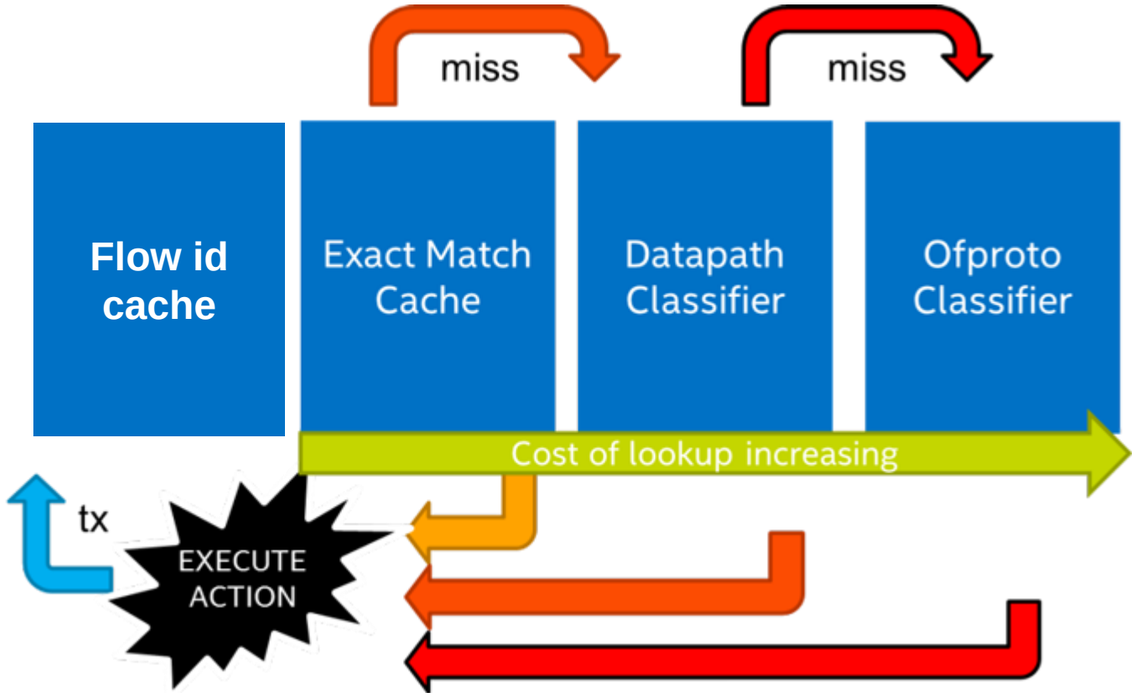
HW classification offload concept

- 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



RFC OVS-DPDK using HW classification offload

- 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



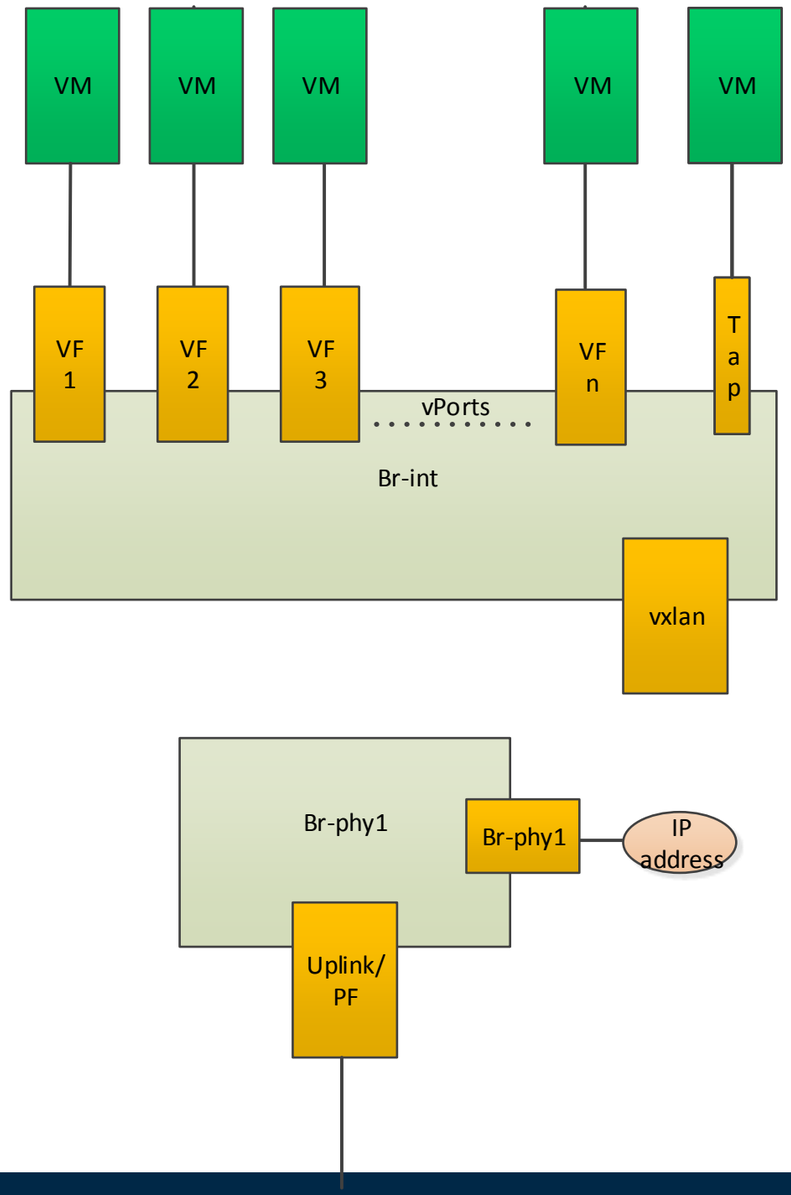
RFC Performance

Case	#flows	Base MPPs	Offload MPPs	improvement
Wire to virtio	1	5.8	8.7	50%
Wire to wire	1	6.9	11.7	70%
Wire to wire	512	4,2	11,2	267%

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

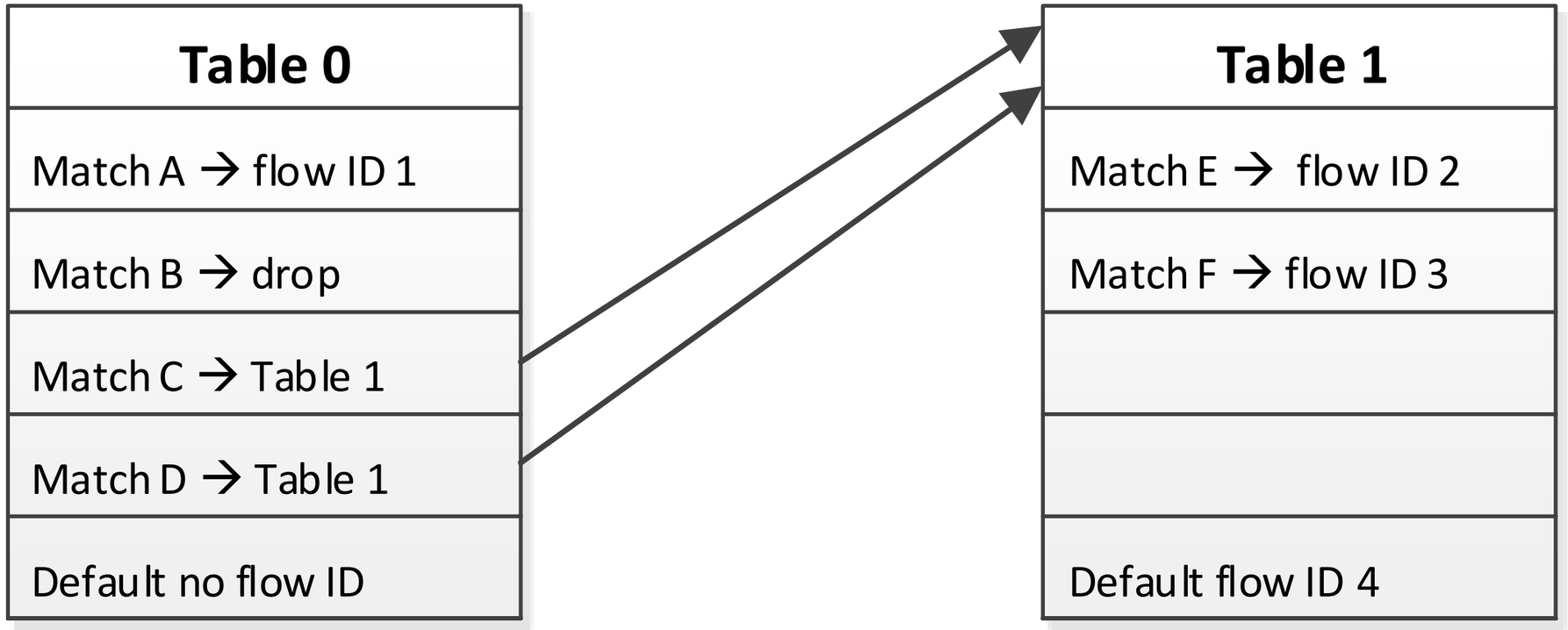
Vxlan in OVS DPDK

- 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



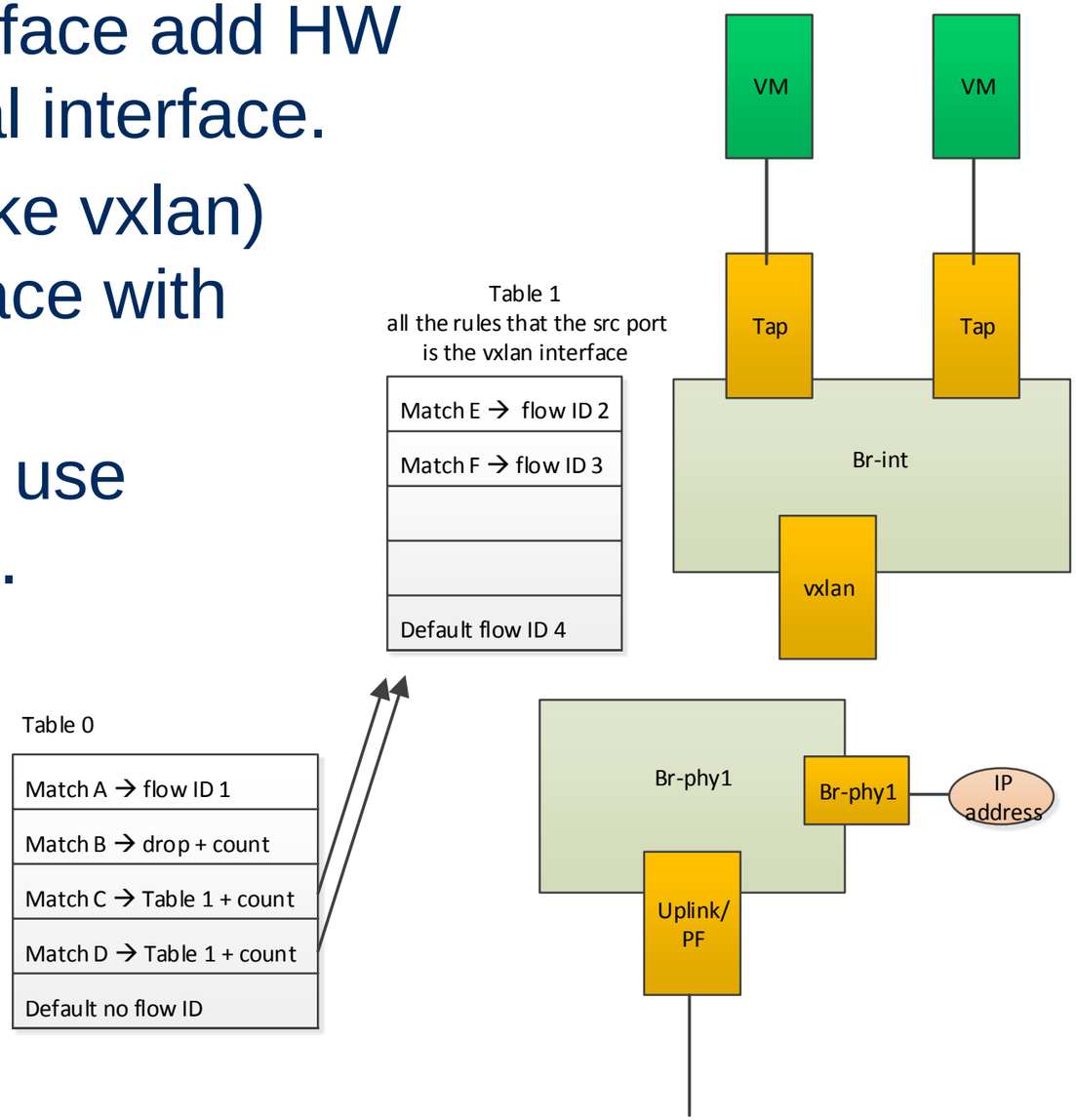
Multi-table

- The action of a rule can be to go to other table.
- It can be use to chain classification



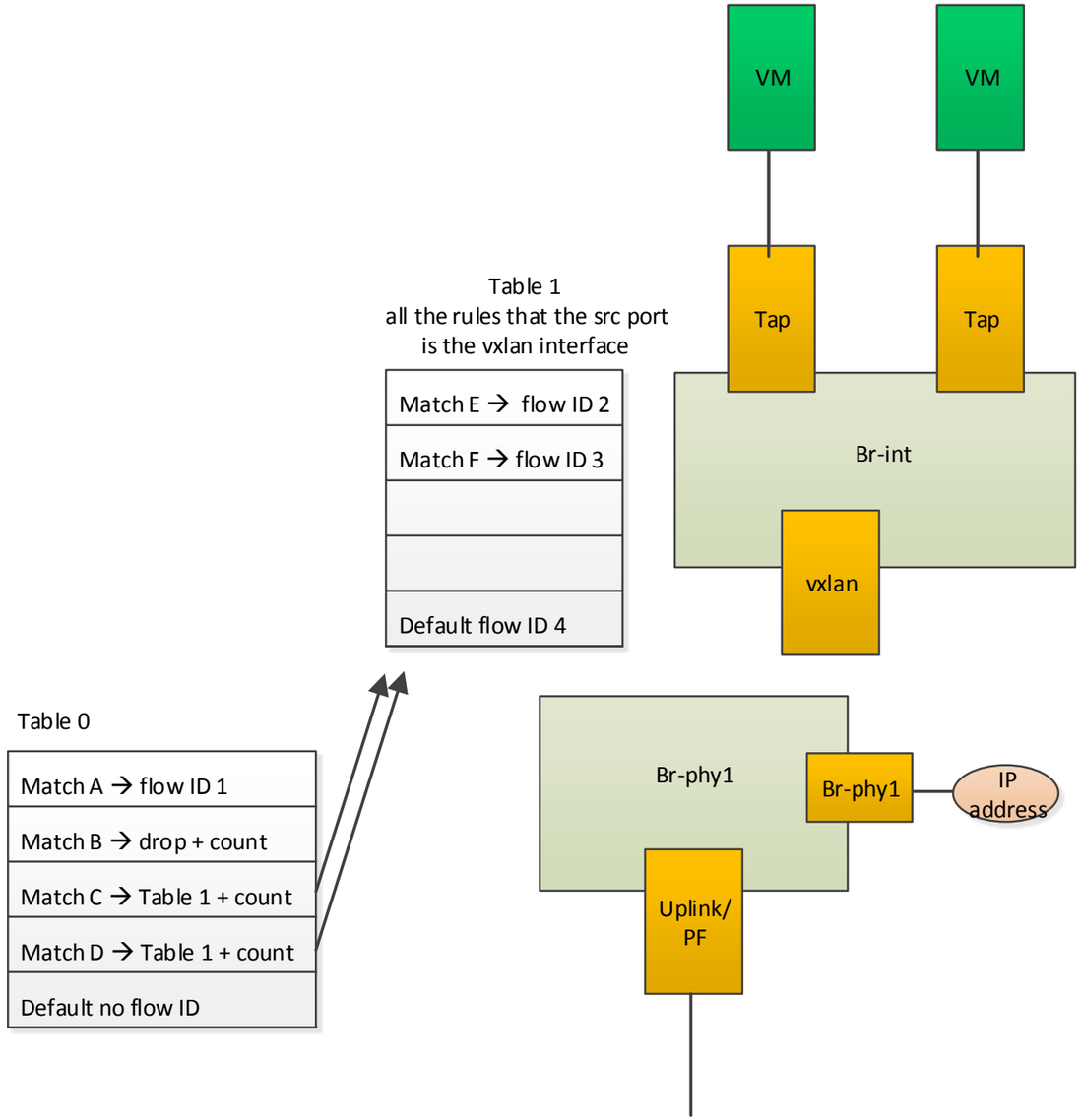
vxlan HW offload concept

- 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 priory) rule to steer the traffic to a Q and not to continue to next group

Thank You

Mellanox TECHNOLOGIES

Connect. Accelerate. Outperform.®