OVS Performance on Steroids - Hardware Acceleration Methodologies

Mellanox TECHNOLOGIES

Connect. Accelerate. Outperform.®

OVS Confernece Nov 2017 Rony Efraim

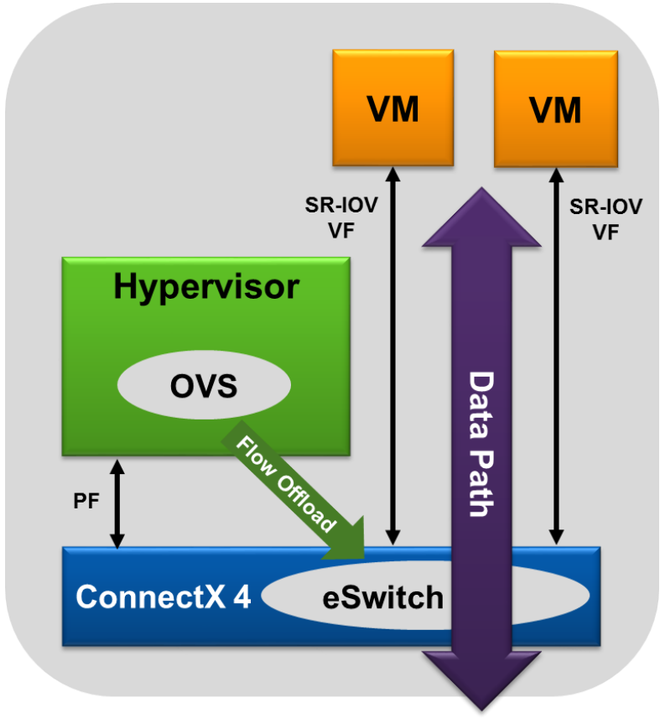
Agenda

- OVS Offload - Accelerated Switch And Packet Processing (ASAP^2)
- Full OVS Offload (ASAP^2 Direct) - SRIOV
  - Software based VS Hardware based
  - OVS support for HW offload
  - OVS HW Offload – ConnectX-5 performance
  - Future work for HW offload
- Partial OVS Offload (ASAP^2 Flex) - DPDK
  - RFC OVS-DPDK using HW classification offload
  - Vxlan in OVS DPDK
  - Multi-table
  - vxlan HW offload concept
- Mellanox OVS Offload Community Work

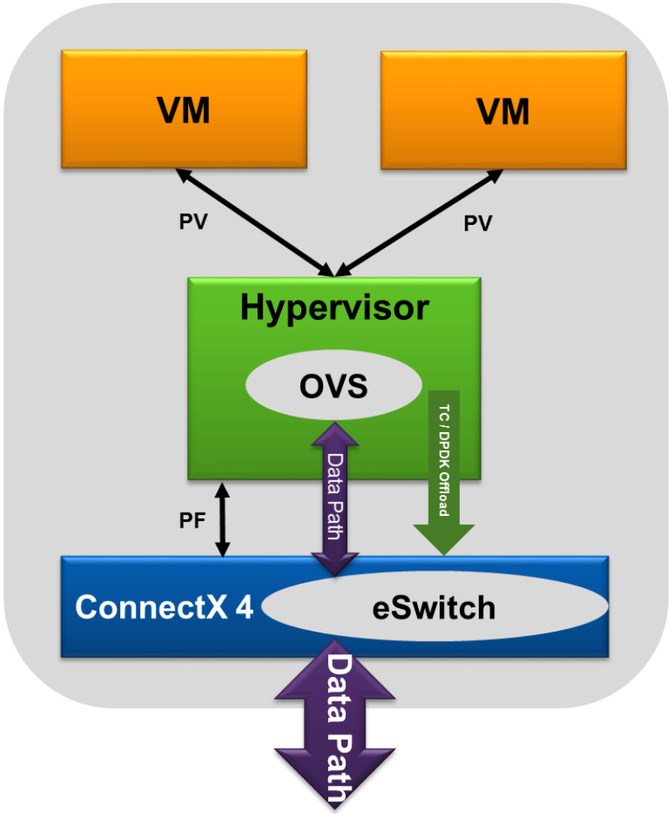
OVS Offload - Accelerated Switch And Packet Processing (ASAP²)

- ASAP² takes advantage of ConnectX-4/5 capability to accelerate \ offload “in host” network stack
- Two main use cases:

ASAP² Direct
Full vSwitch offload (SR-IOV)



ASAP² Flex
vSwitch acceleration

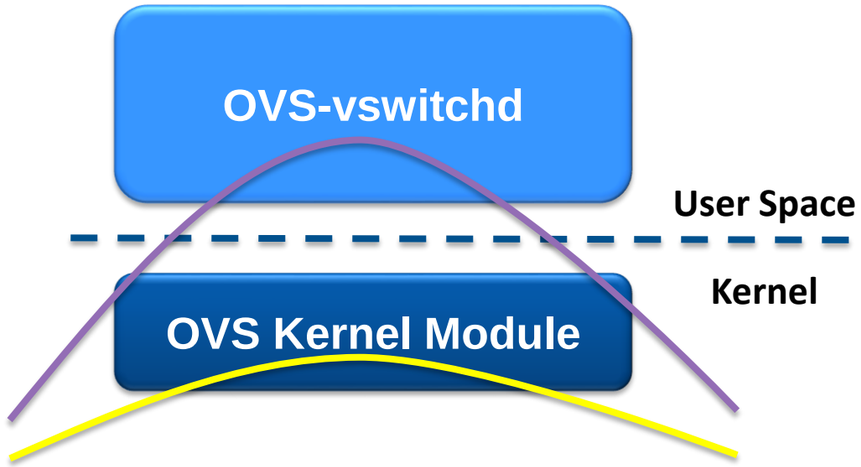


ASAP²-Direct

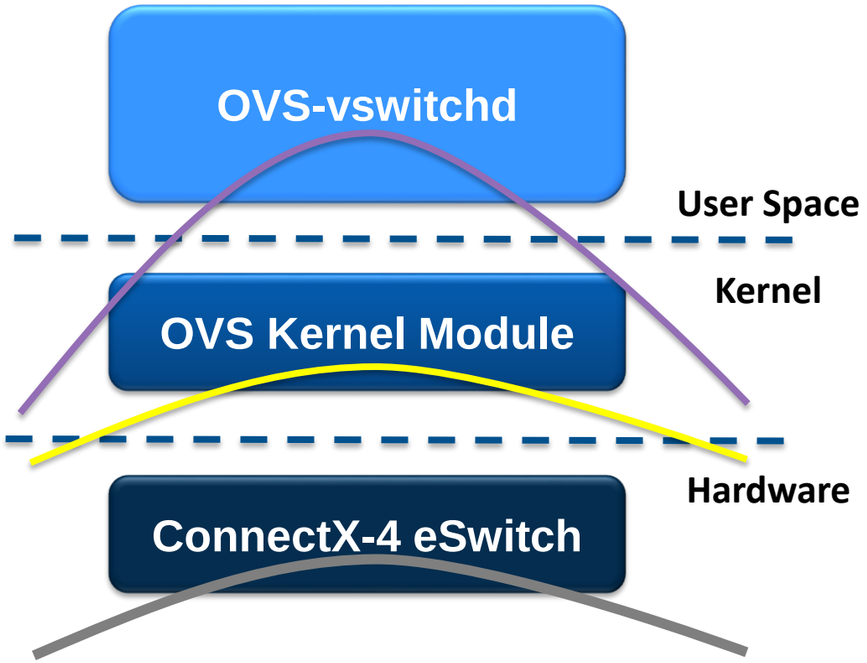
Full Virtual Switch Offload (SRIOV)

Software based Vs. Hardware based

Traditional Model: All Software
High Latency, Low Bandwidth, CPU Intensive



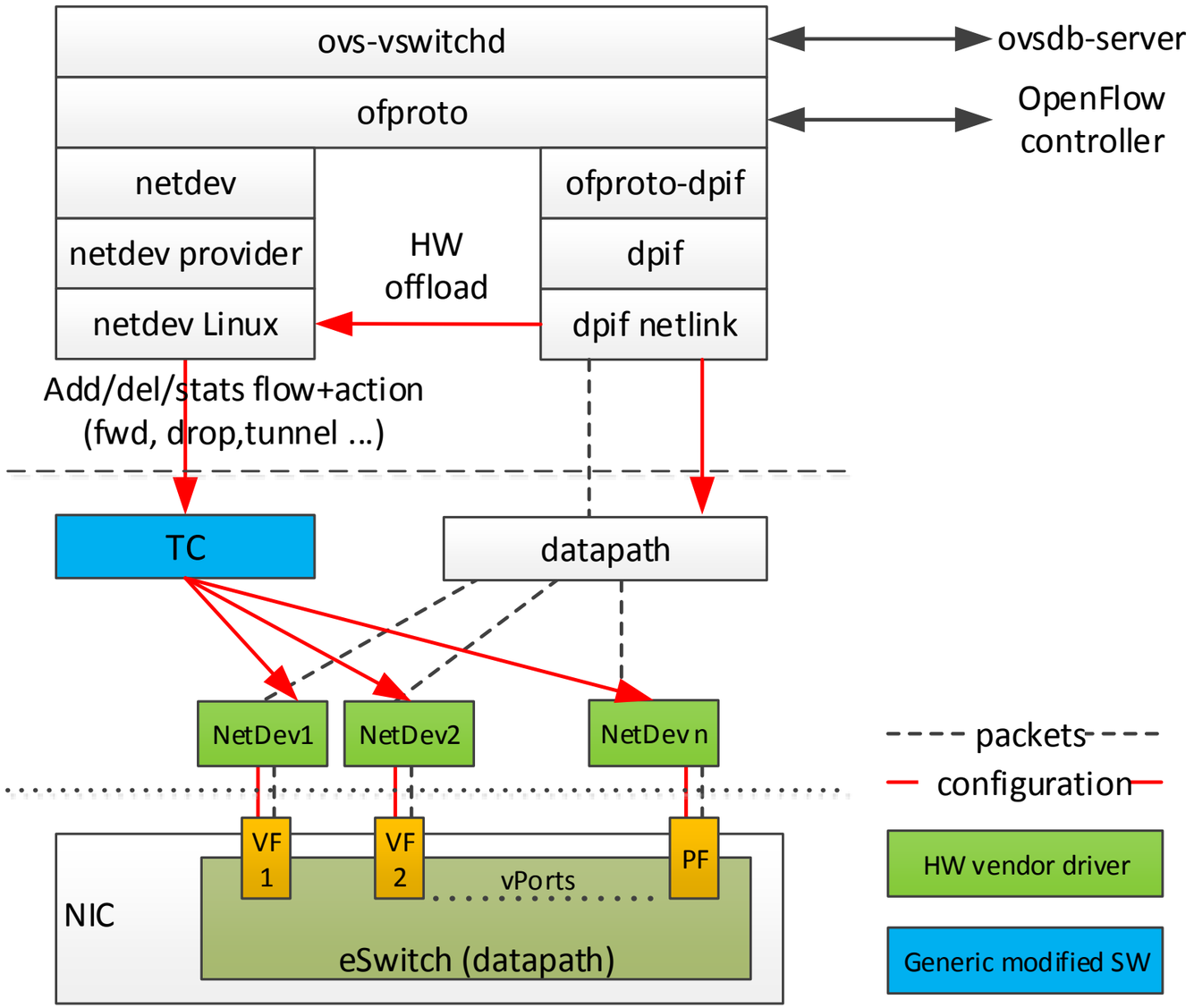
ConnectX-4: Hardware Offload
Low Latency, High Bandwidth, Efficient CPU



First flow packet Fallback FRWD path HW forwarded Packets

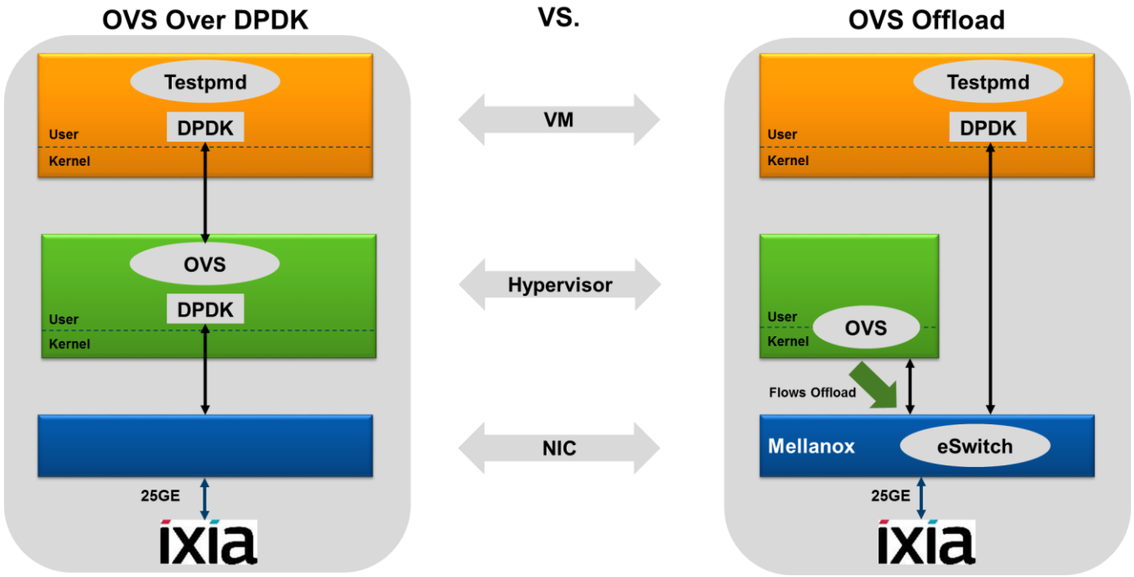
OVS support for HW offload

- Changes are made only in the OVS user space code.
- HW offload of flow using TC flower.
- Packets forwarded by the kernel datapath are transmitted on the representors and forwarded by the e-switch to the respective VF or to the wire

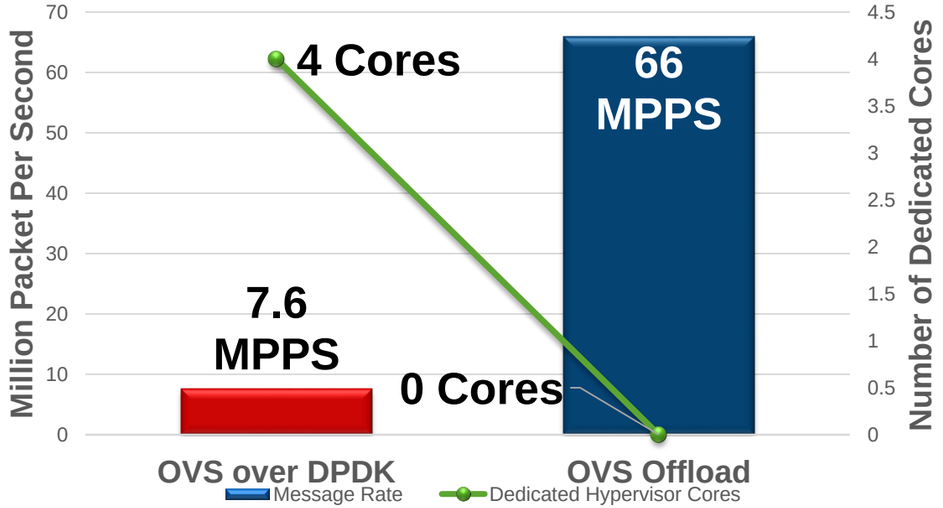


OVS HW Offload – ConnectX-5 performance

Test	ASAP2 Direct	OVS DPDK	Benefit
1 Flow VXLAN	66M PPS	7.6M PPS (VLAN)	8.6X
60K flows VXLAN	19.8M PPS	1.9M PPS	10.4X



- ConnectX-5 provide significant performance boost
  - Without adding CPU resources



Future work for HW offload

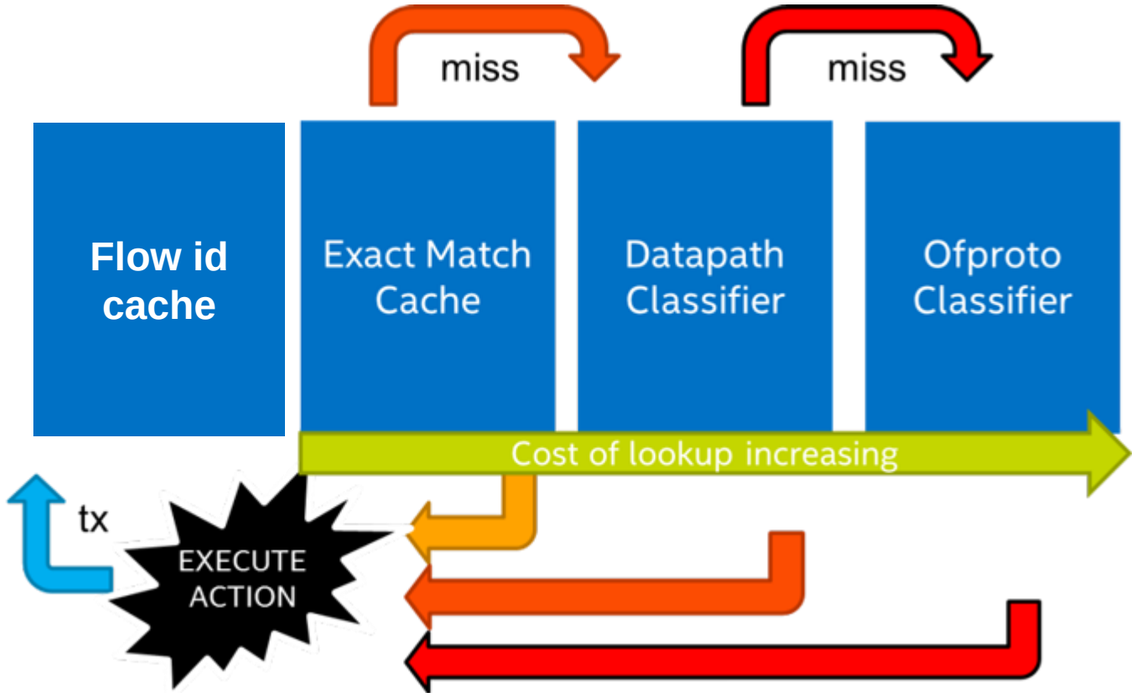
- Table offload to support recirculate
- Connection tracking
- LAG (bonding) for SRIOV
- VF live migration

ASAP² Flex

HW accelerate OVS-DPDK

RFC OVS-DPDK using HW classification offload

- For every datapath rule we add a rte_flow with flow Id
- The flow id cache contains mega flow rules
- When packet is received with flow id, no need to classify the packet to get the rule



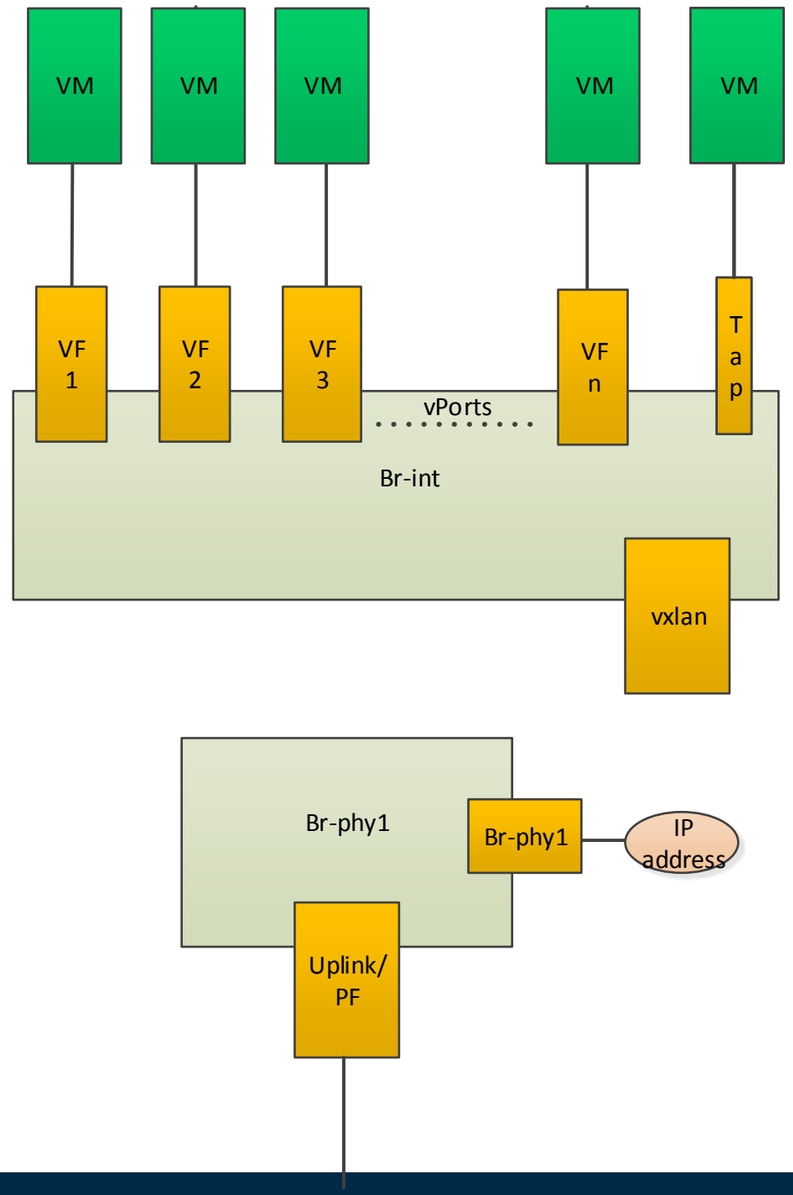
RFC Performance

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.

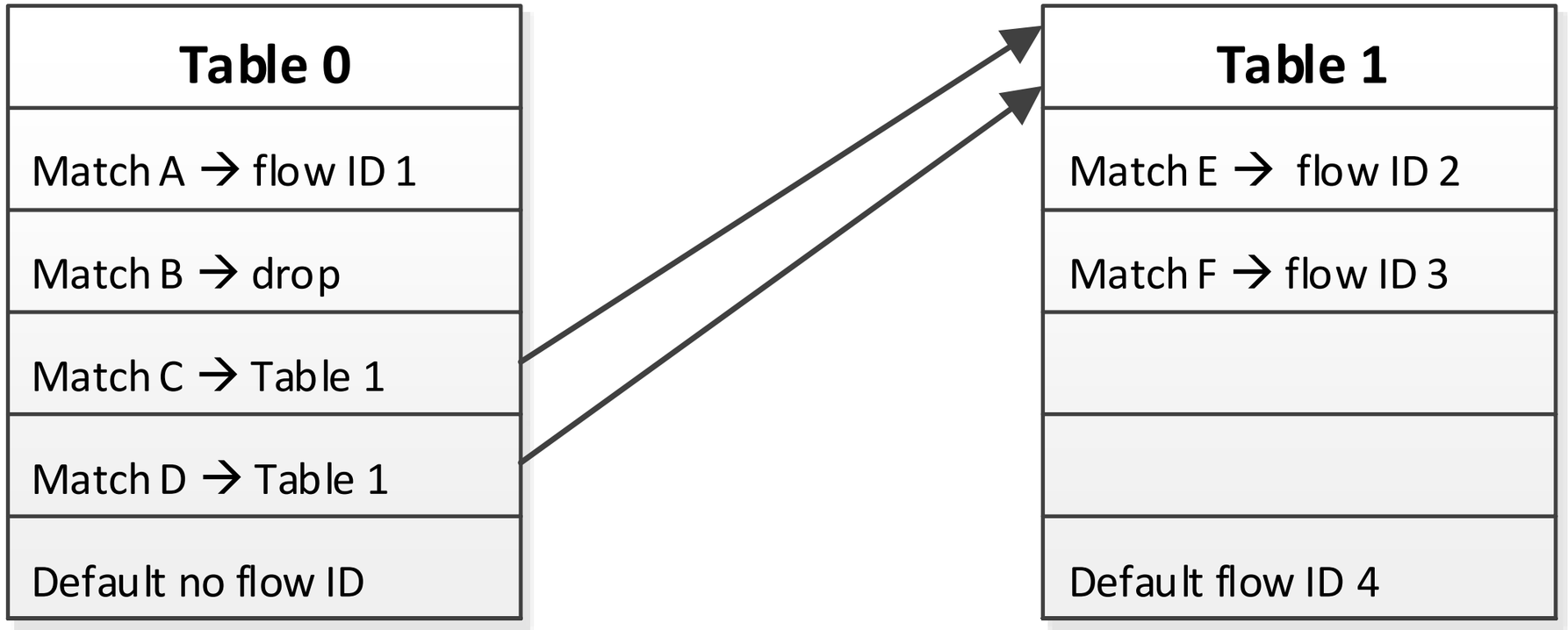
Vxlan in OVS DPDK

- There are 2 level of switches that are cascaded
- The HW classification accelerates only the lower switch (br-phys1)
- br-phy1 is a kernel interface for vxlan
- The OVS datapath is required to classify the inner packet



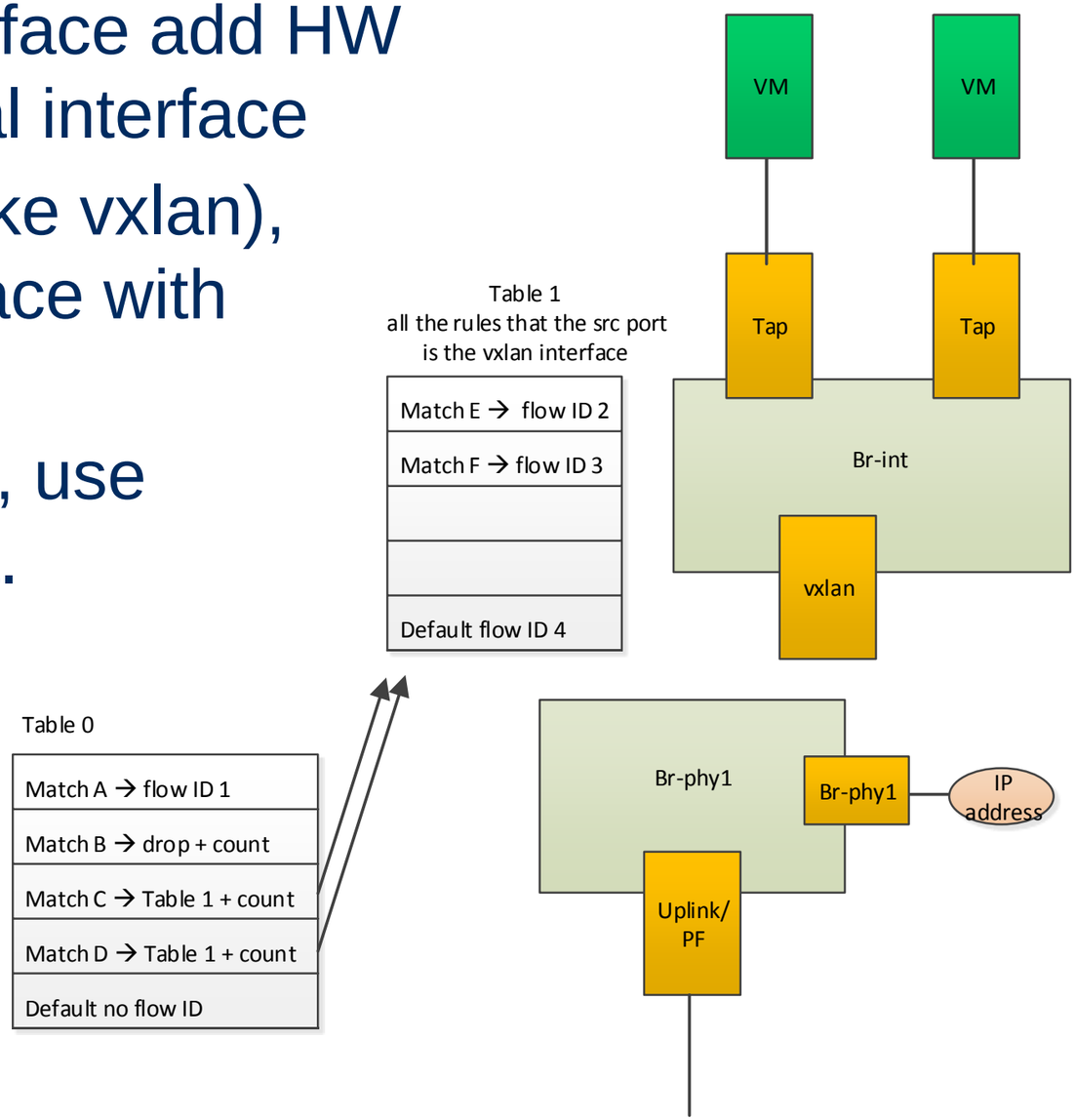
Multi-table

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



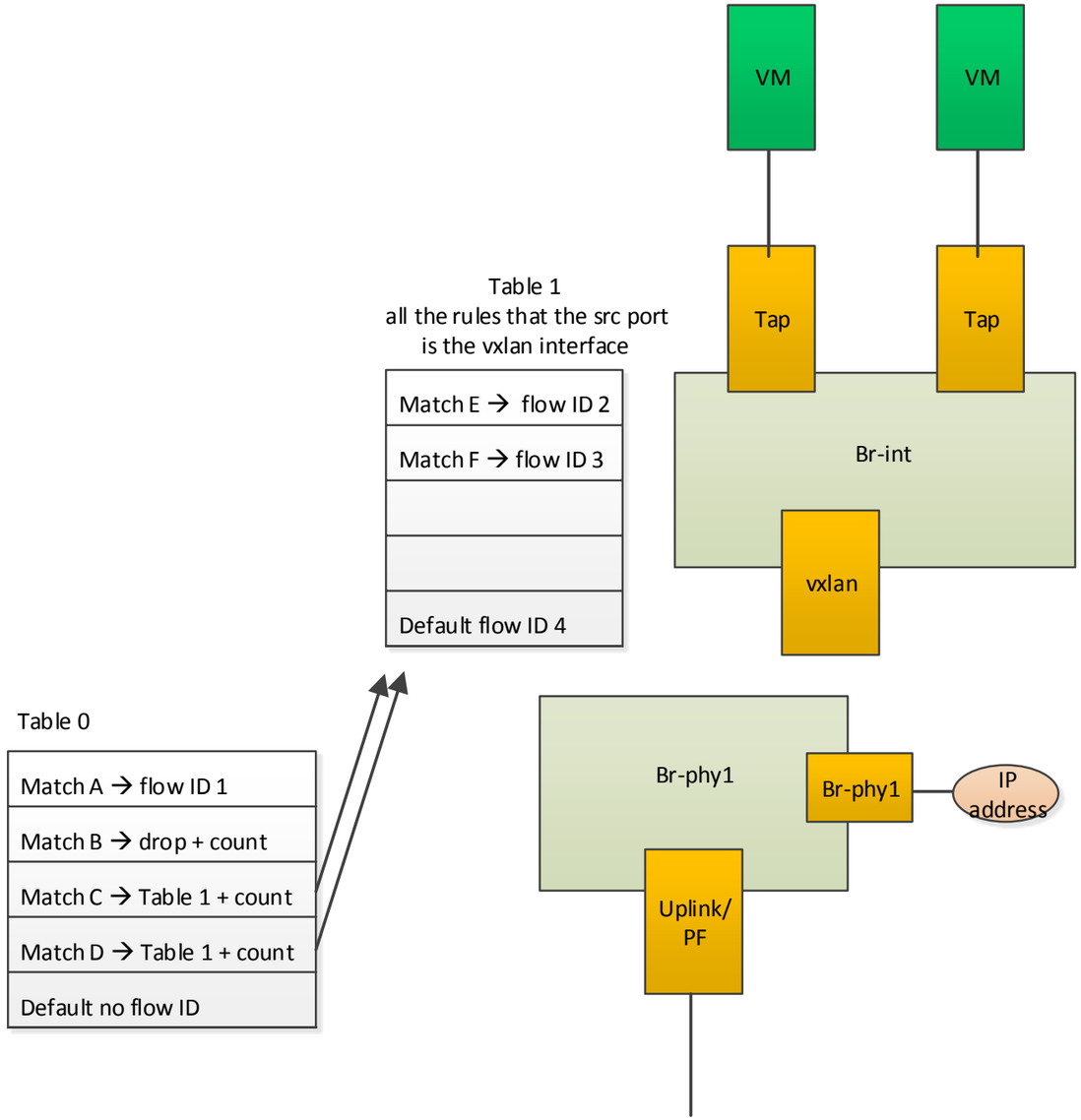
VxLAN HW offload concept

- 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 a physical port and is classified according to the outer and the inner headers



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



Mellanox OVS Offload Community Work

- Linux Kernel :
  Using SR-IOV offloads with Open-vSwitch – netdev conf 1.2
  https://netdevconf.org/1.2/papers/efraim-gerlitz-sriov-ovs-final.pdf
- Open vSwitch :
  [PATCH V11 00/33] Introducing HW offload support for openvswitch
  https://mail.openvswitch.org/pipermail/ovs-dev/2017-June/333957.htm
- Open Stack
  Os-vif : https://review.openstack.org/#/c/398277/
  Nova : https://review.openstack.org/#/c/398265
  Neutron : https://review.openstack.org/#/c/275616
- DPDK
  RTE_Flow API
  OVS-DPDK RFC:
  https://www.mail-archive.com/ovs-dev@openvswitch.org/msg12562.html

ASAP^2 Mellanox team

- Roi Dayan
- Paul Blakey
- Hadar Hen Zion
- Mark Bloch
- Or Gerlitz
- Natali Shechtman
- Natan Oppenheimer
- Oded Shanoon
- Olga Shern
- Yuanhan Liu
- Moshe Levi
- Rabie Loulou
- Rony Efraim
- Shahar Klein
- Shani Michaeli
- Tal Anker
- Haggai Eran
- Ilya Lesokhin
- Lior Narkis
- Vlad Buslov
- Chris Mi

Thank You

Mellanox TECHNOLOGIES

Connect. Accelerate. Outperform.®