Using SR-IOV offloads with Open-vSwitch and similar applications

Rony Efraim, Or Gerlitz

Netdev **conferences** 1.2 Oct 2016

Mellanox TECHNOLOGIES

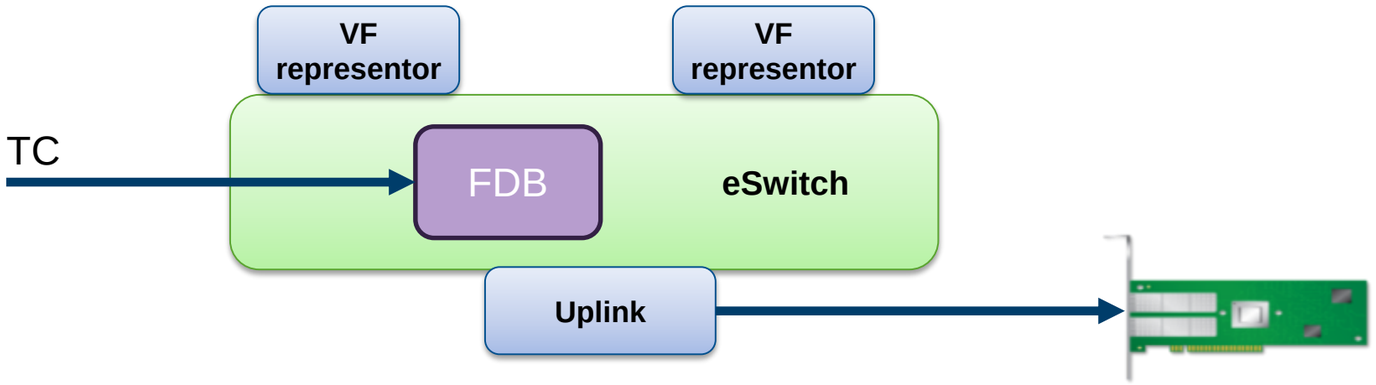
Connect. Accelerate. Outperform.™

Agenda

- Solution building blocks
- Use the software datapath (kernel) as a slow path
- Flow aging & counters
- Policy considerations
- Working example of HW offloads with Open-VSwitch (OVS)
- Nested Containers use case

Solution building blocks

- vPort - VF representors
  - send /receive traffic from the VF
  - vPort counters
  - control the link state
- eSwitch FDB - TC to config flow based forwarding rules and ACL
- Aging - TC to read flow counters
- Tunnel - tunnel netdev and TC in order to HW offload rules



Use the software datapath (kernel) as a slow path

- By default a miss rule is installed in the HW
- Packets not belonging to any offloaded flow hit the miss rule
- Missed packet is received into the host OS from the representor of that port (VF or uplink) according to the real vPort thy ware received on
- The packet will show up in the software data-path and would experience a miss there too. Once this happens, the packet is sent to the switch management software
- If the decision is to offload the flow, a TC call will be made to configure that into the HW through the respective representor
- The original packet will be re-injected to the e-switch on the egress port representor which is dictated from the new flow and hence will show up in the correct destination

Flow aging & counters

- Data driven switch population rules are based on traffic
- Remove the rule when there is no traffic that matches that rule
- When flow is offloaded to HW, the software counters will be always 0
- The switch software needs to read HW counters for the offloaded flows
- The mechanism to read the HW traffic counters and last used time of flow rules is TC
- The switch software polls the counters and last used of all the rules
- Removes the rules that were not been used for long time

Policy considerations

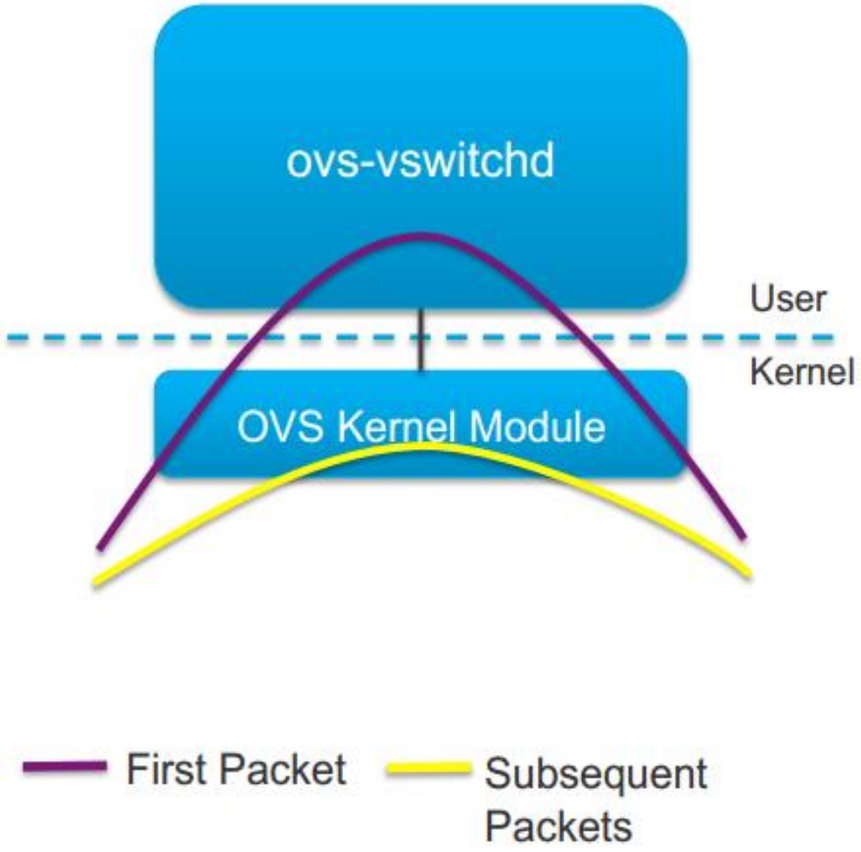
- HW acceleration scheme should support the ability for the system administrator to dictate which flows get offloaded and which not

- The reasons:
  - pure-policy (i.e. to differentiate between customers/VMs/etc.)
  - complex processing of flows that require CPU handling (e.g. flows that requires statefull inspection/DPI)

- To achieve that, some policy module is needed, in order to decide which flows get offloaded and which not

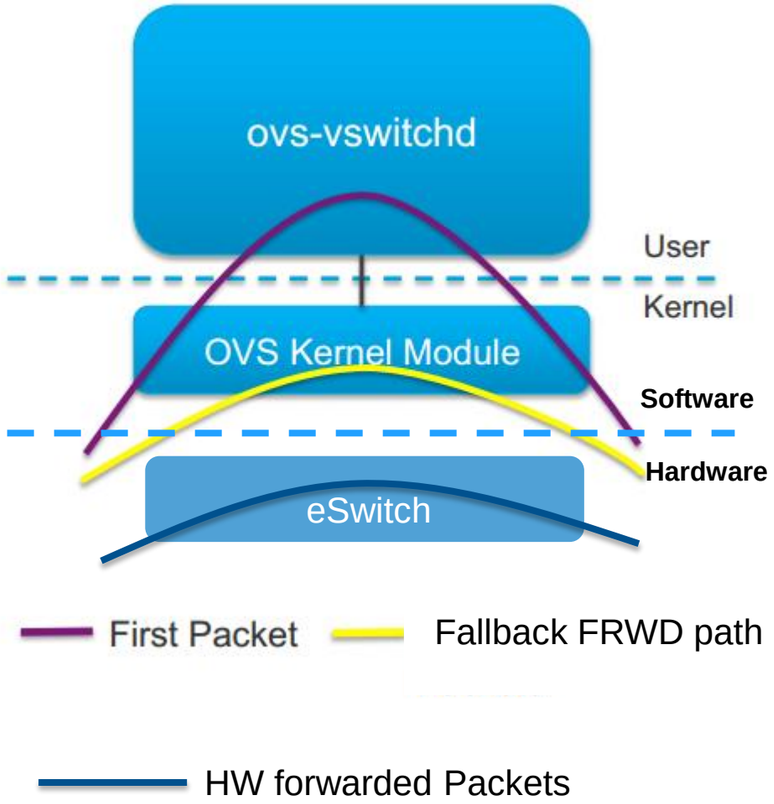
example of HW offloads with Open-VSwitch (OVS)

- One of the dominant virtual switches is OVS (Open Virtual Switch)
- Forwarding
  - Flow-based forwarding
  - Decision about how to process a packet is made in user space
  - First packet of a new flow is directed to ovs-vswitchd, following packets hit cached entry in kernel
- OVS Overview
  - http://openvswitch.org/slides/OpenStack-131107.pdf



OVS Offload – Solution: Adding the Hardware Layer to the Forwarding Plane

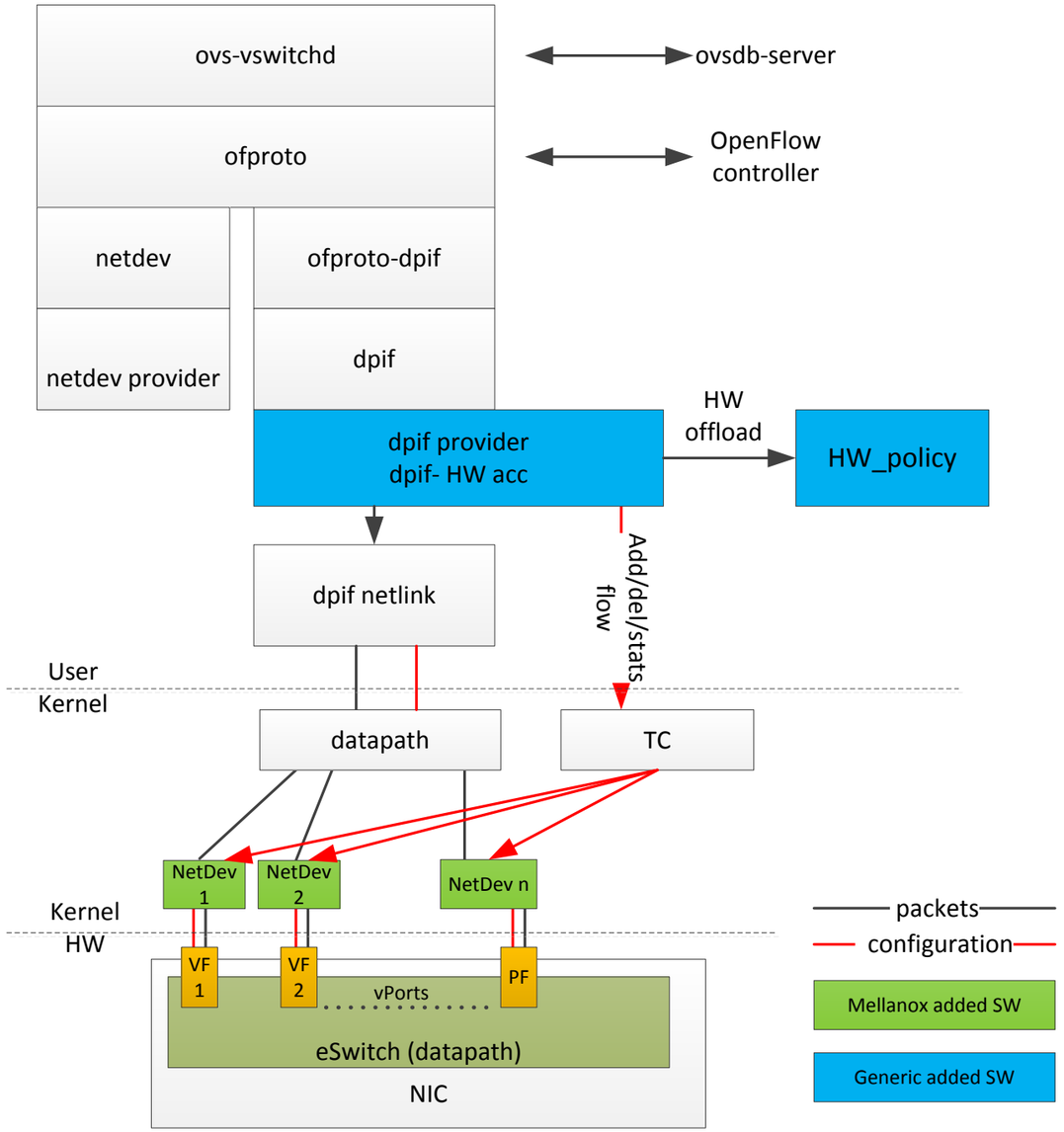
- The NIC Embedded Switch is layered below the kernel datapath
- The Embedded Switch is the first to ‘see’ all packets
- New flow (‘miss’ action) is directed to OVS kernel module
  - Miss in kernel will forward the packet to user space as before
- Decision if to offload the new flow to HW is done by “Offload Policer” based on device capabilities
- Following packets of flow are forwarded by eSwitch -- if offloaded



Retain the “first packet” concept (slow path) while enabling the “fast-est” path – via the HW switch by installing the proper flows

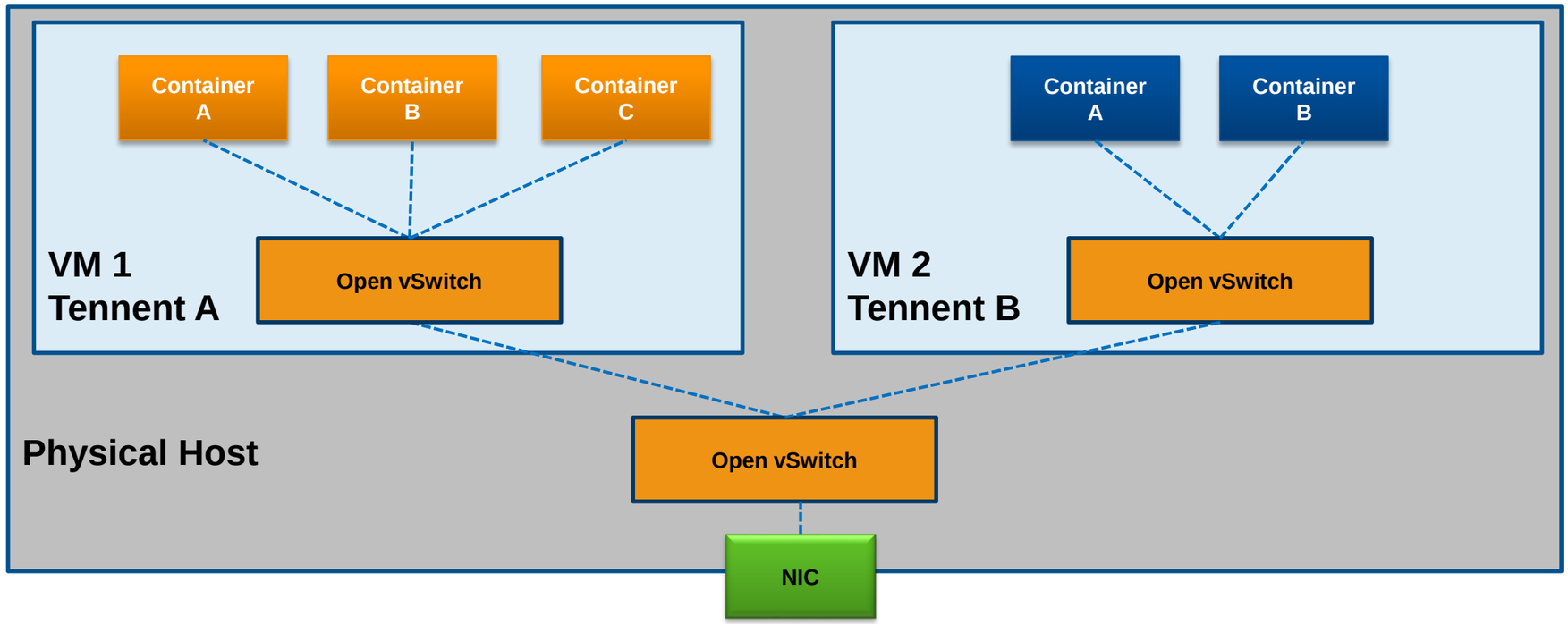
OVS model to support policy based DPIF HW offloading

- Changes are made only in the OVS user space code (No OVS kernel datapath module changes)
- Plugin a new DPIF module dpif-HW-acc
  - Add "*HW_offload_test_...*" APIs
    - All the Policy code in a specific SW module
    - Policy can decide on "HW ONLY/ NO OFFLOAD/ SPLIT"
  - HW offload flow by setting a HW only ("skip sw") TC rule
  - Aging of old flows is done by polling all the rules every few seconds. The offloaded HW rules are polled too through TC and removed according to aging policy of OVS
  - 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
  - Link to RFC: http://openvswitch.org/pipermail/dev/2016-September/079952.html



Nested Containers use case

- Containers in VMs
- Leverage existing Virtualization / Cloud infrastructure
- Challenge: Two layers virtual switching -> significant performance impact




time for questions

Thank You

Mellanox TECHNOLOGIES

Connect. Accelerate. Outperform.™