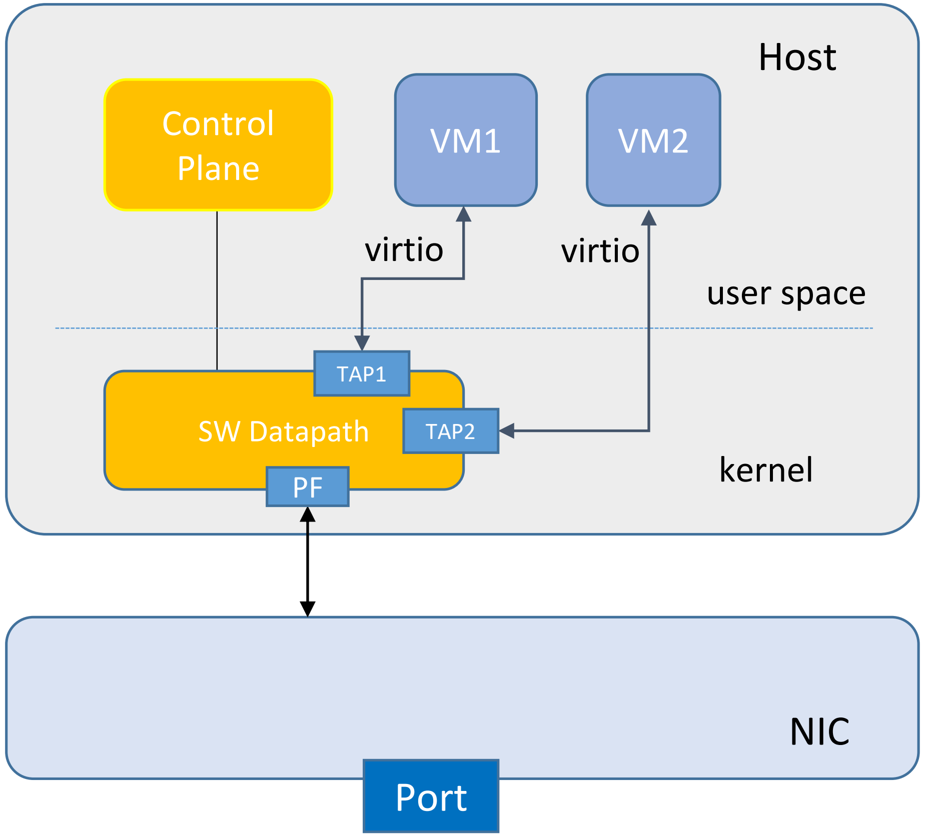
vSwitch Acceleration with Hardware Offloading

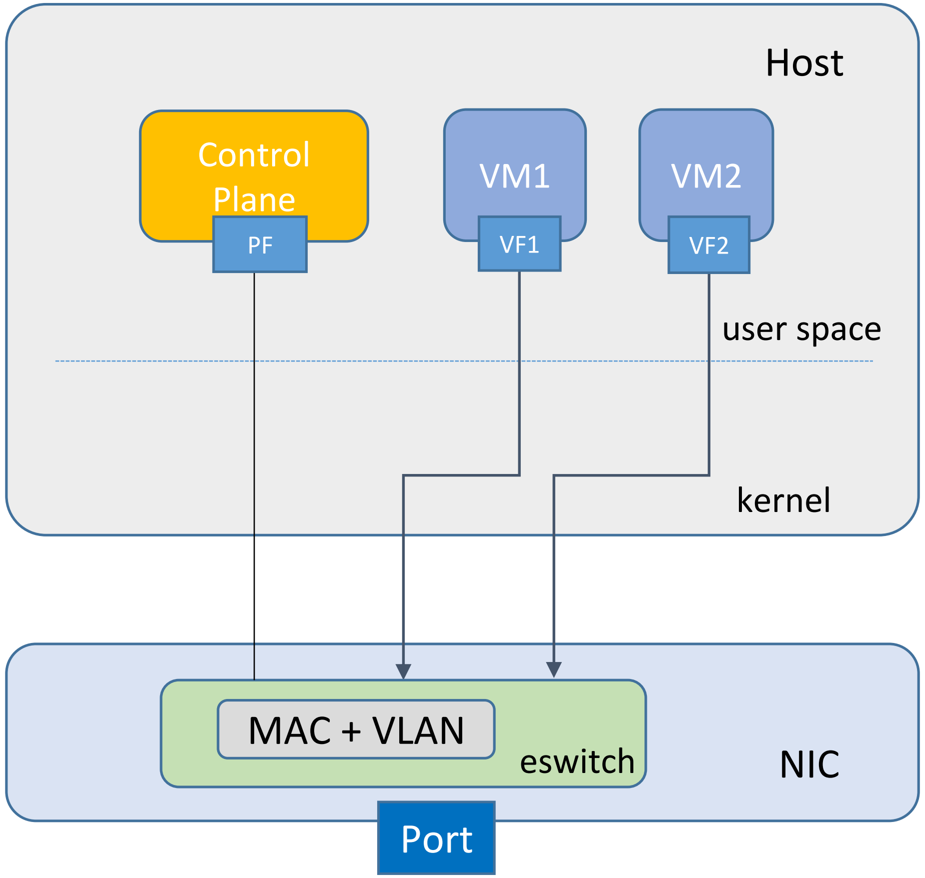
CHEN ZHIHUI

JUNE 2018

Current Network Solution for Virtualization

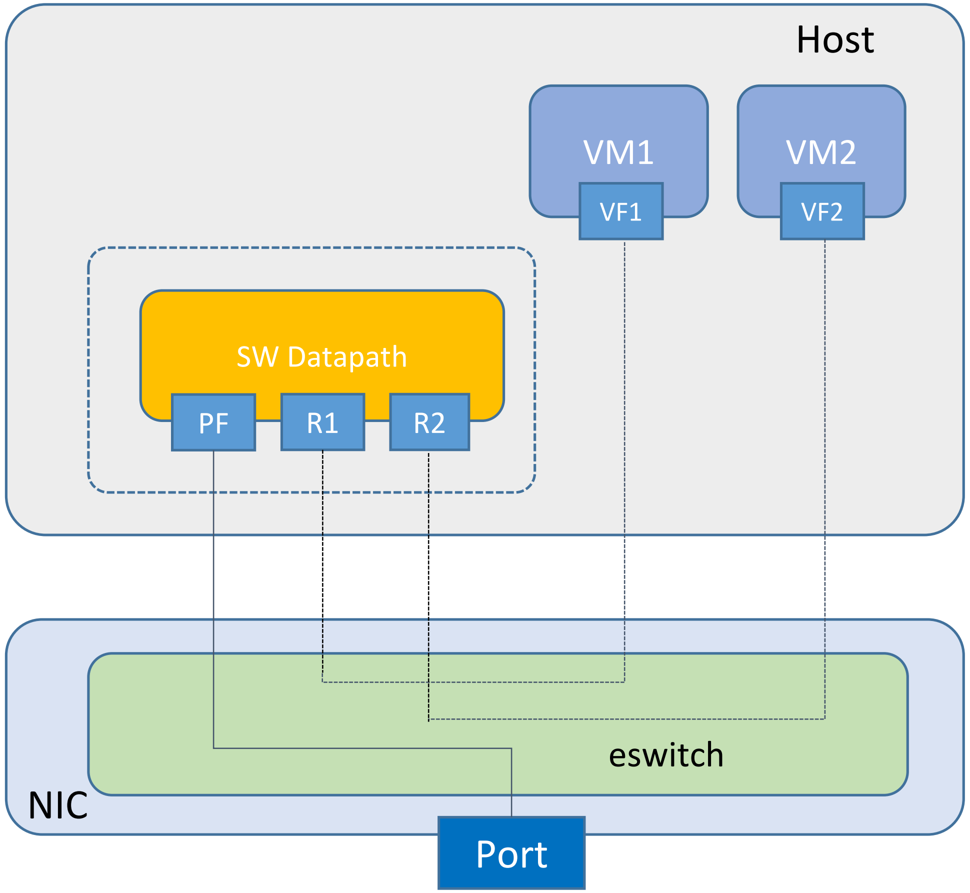


Software Solution



SRIOV Solution

VF Representor for Virtualization



- VF Representor
  - Net Device modeling of eswitch port and exposed through PF driver.
  - VF and its representor works like Linux veth pair
  - Flow configuration (add/remove)
  - Works under switchdev mode
- Access from both kernel and DPDK
  - Multi Queue (RSS/TSO/CSUM)
  - Attach/Detach in DPDK
  - Multiple DPDK instances over VF representor
- With VF representor, vSwitch can work with SRIOV together and reduce CPU% consumed by virtio.
- Disadvantages:
  - 3x PCIe access for traffic from VM to wire and vice versa, PCIe can become a bottleneck for throughput.
  - Need vendor specific driver in VM.

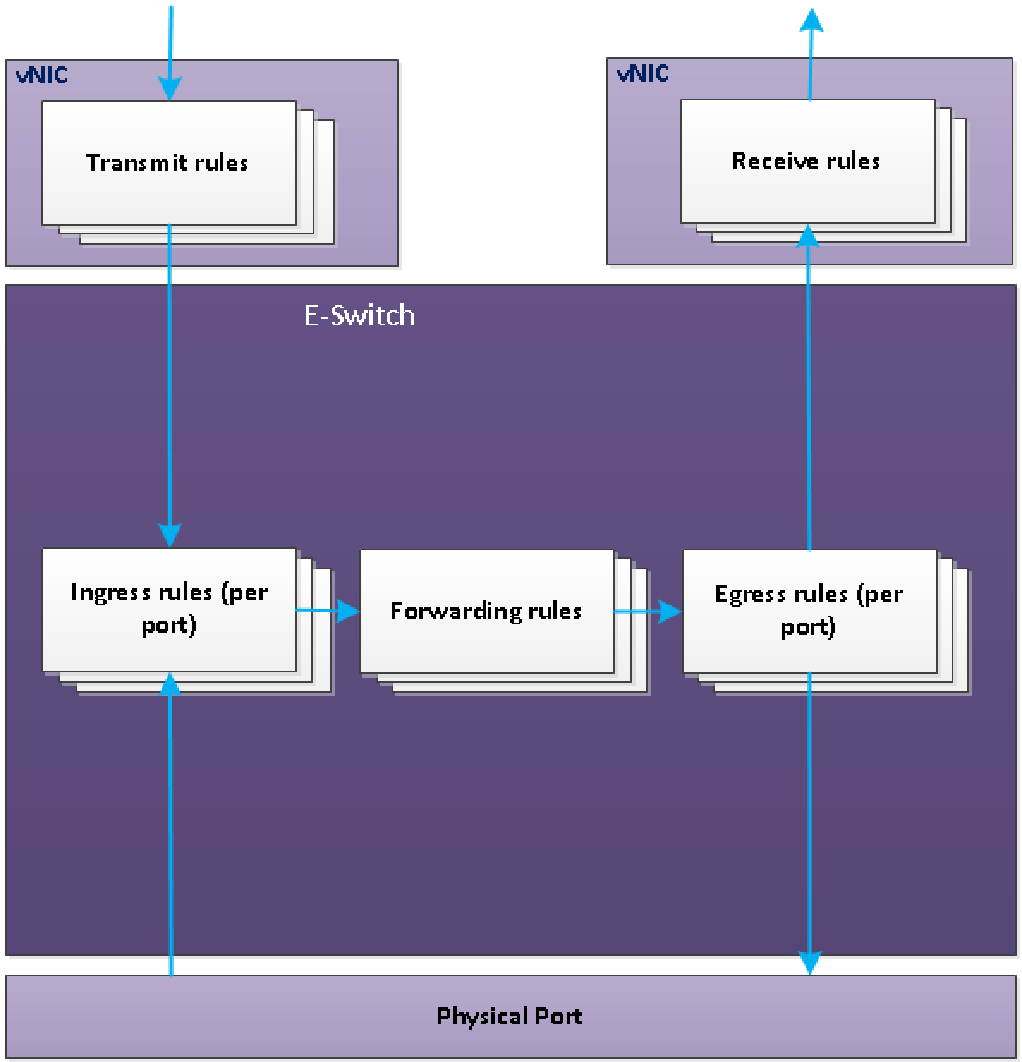
Flow Table with Mellanox Adapter

- Key match fields
  - Ethernet
  - IP(v4 /v6)
  - TCP/UDP
  - Inner packet for Overlay
  - VNI
- Flexible fields extraction by “Flexparse”

- Action
  - Forwarding
  - Drop
  - Counter
  - Encap/Decap
  - Flow ID
  - Header rewrite
  - ……

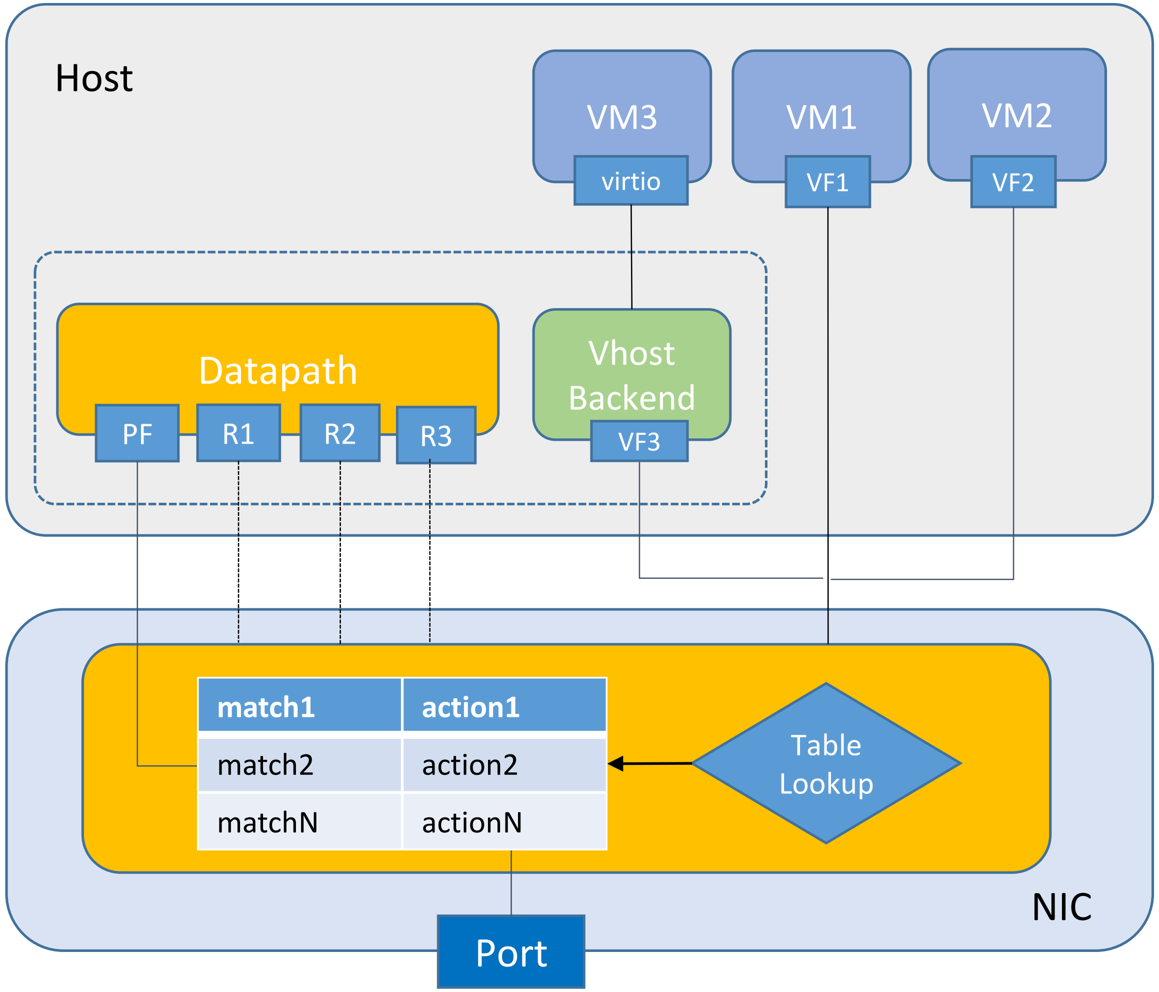
Classification					Action
SRC MAC =	Dest MAC =	SRC IP =	Dest IP =	Protocol =	Counter
SRC MAC =	Dest MAC =	SRC IP =	Dest IP =	Protocol =	Another rule

Classification					Action
VLAN tag =	Tunneling type	Inner packet SRC IP =	Inner packet Dest IP =	Inner packet Protocol =	Header re-write
VLAN tag =	Tunneling type	Inner packet SRC IP =	Inner packet Dest IP =	Inner packet Protocol =	Meta Data
VLAN tag =	Tunneling type	Inner packet SRC IP =	Inner packet Dest IP =	Inner packet Protocol =	Flow-ID Tag



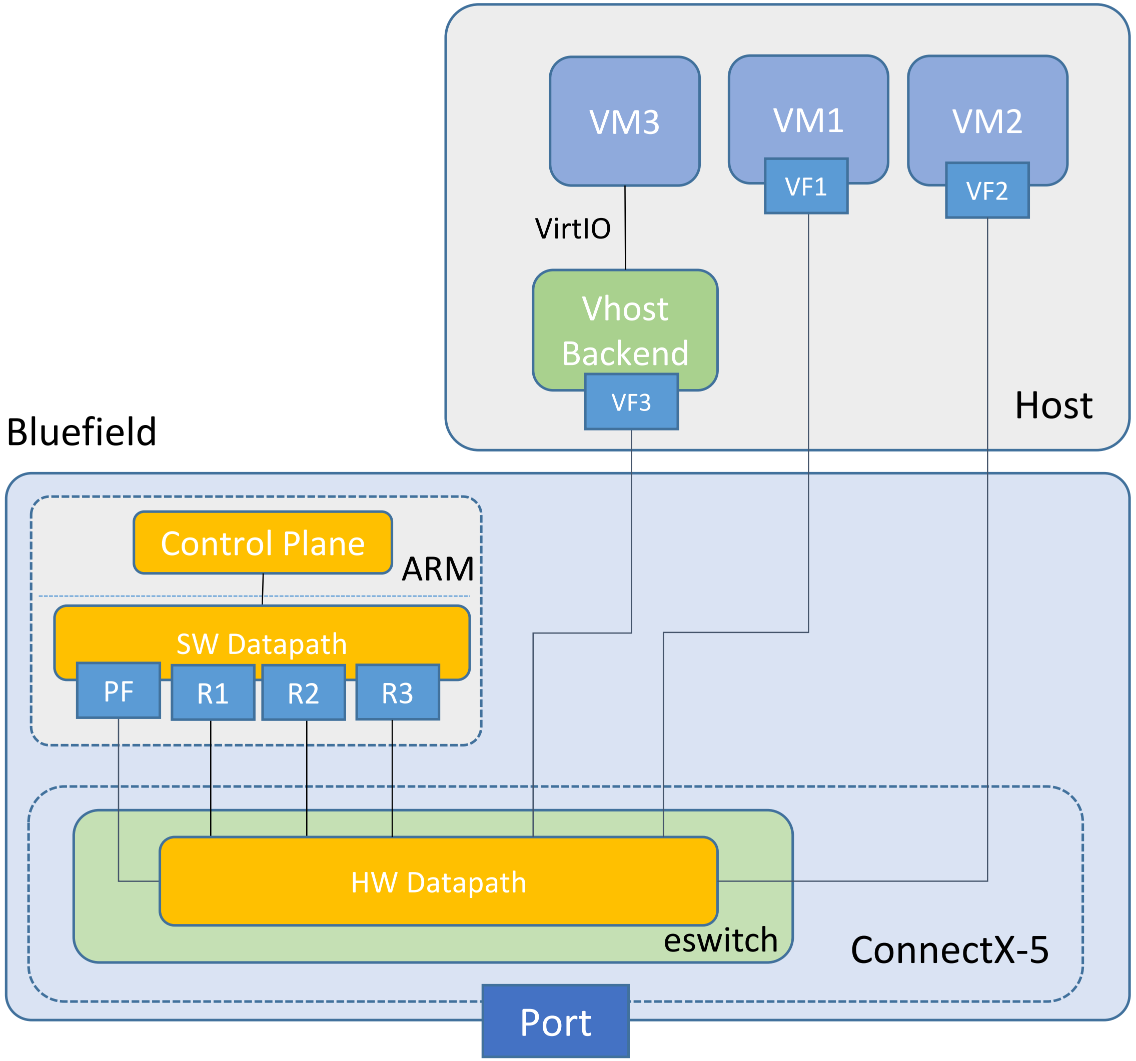
vSwitch Design with Hardware Offloading

- HW datapath on eswitch through configuring flow table.
    - TC Flower
    - DPDK RTE_FLOW
- Software datapath handle ‘the first packet’ and unsupported flows through VF representor.
- Support both SRIOV and VirtIO
    - Direct path to VM for SRIOV
    - Optimized vhost backend for virtio acceleration
        - **TX**: Forward packet to HW with meta data.
        - **RX**: Receive packet from VF with Flow ID which can identify destination VM.
- Rules management
    - Add/delete/Query
    - Aging
    - Batch operations



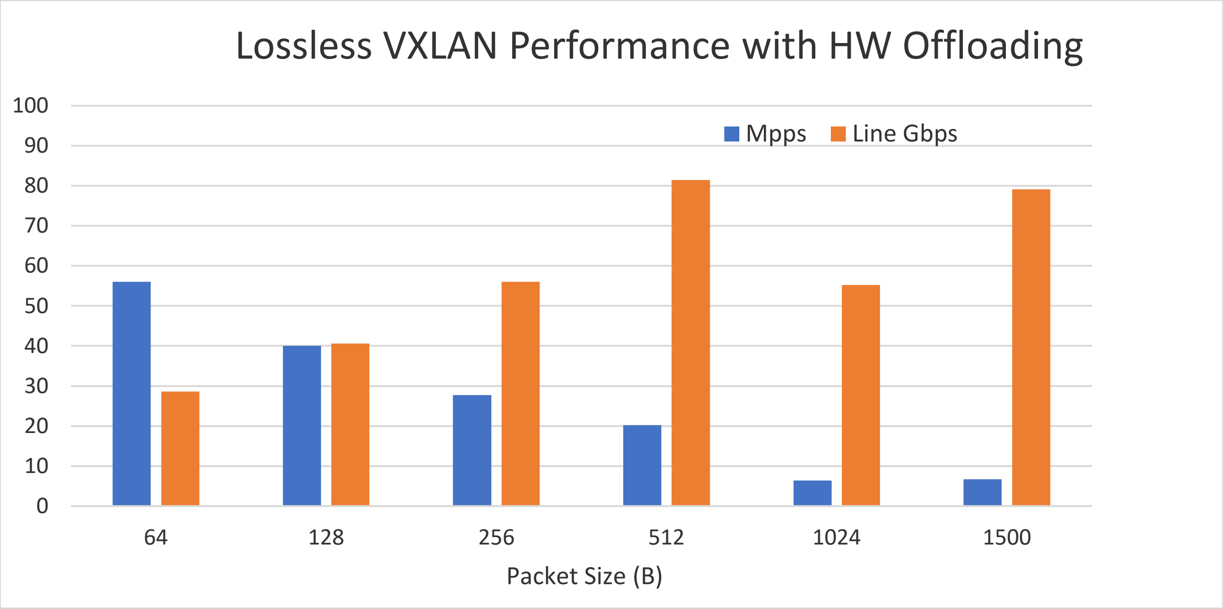
match1	action1
match2	action2
matchN	actionN

vSwitch with Hardware Offloading for SmartNIC



- Control plane and SW Datapath on ARM
- HW Datapath on NIC
- Both SRIOV and VIRTIO interface to VM
- Advantages
  - Support bare-metal cloud
  - Separation of computing domain and networking domain, all host resources (core and memory) can be used for VMs.
  - Efficient
- Disadvantages
  - Higher cost and power
  - Two management domain.

Hardware Offloading Performance with SRIOV



System Configuration:

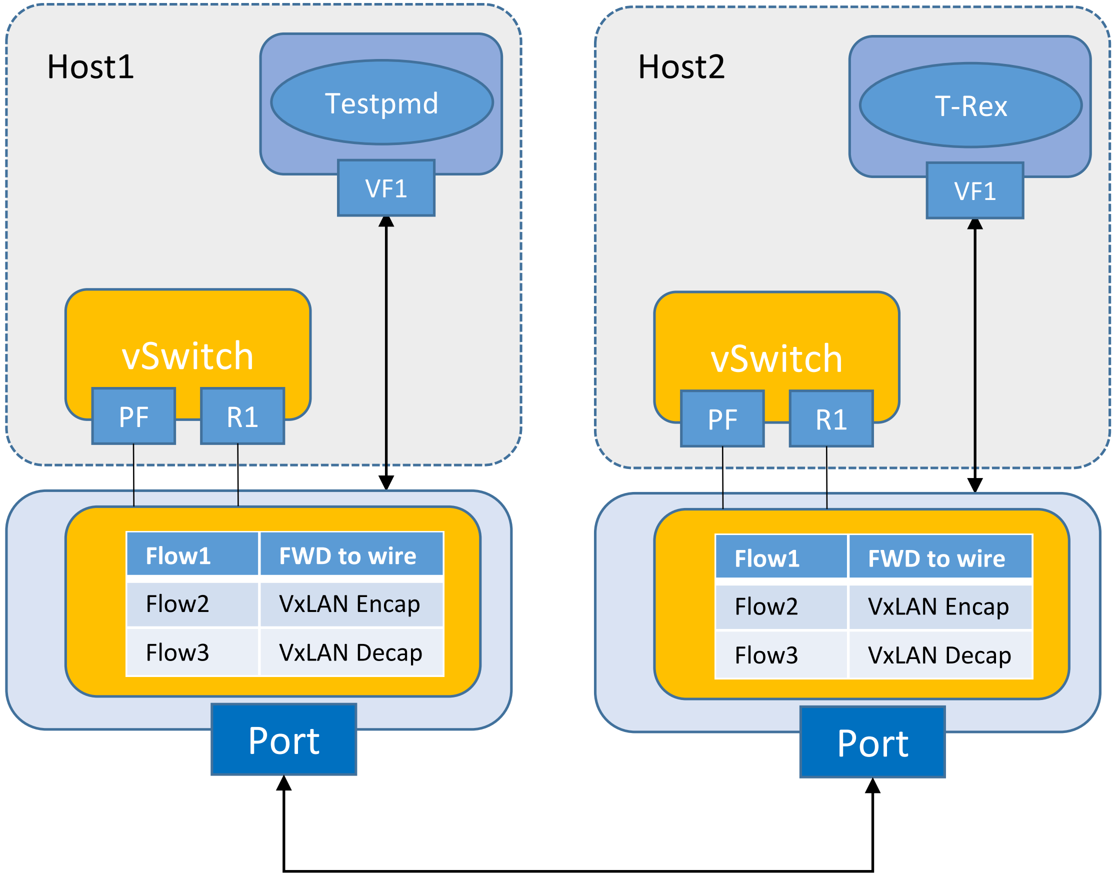
(1) E5-2667 V3 @ 3.20GHz

(2) Mellanox 100G ConnectX-5 NIC

(3) RHEL7.5 Host and Guest

(4) Vxlan Encap/Decap on NIC

(5) T-Rex and Testpmd run in VMs



Flow1	FWD to wire
Flow2	VxLAN Encap
Flow3	VxLAN Decap

Flow1	FWD to wire
Flow2	VxLAN Encap
Flow3	VxLAN Decap

Hardware Offloading Performance with VirtIO

Test Case\Cores	1	2	4
VM->HV->wire	9.98	18.3	36.4
VM->HW->VxLAN Encap-wire	9.95	18.3	36.2
Wire->HV->VM	13.4	23.3	46.3
Wire->VxLAN Decap->HW>VM	13.5	25.0	45.8

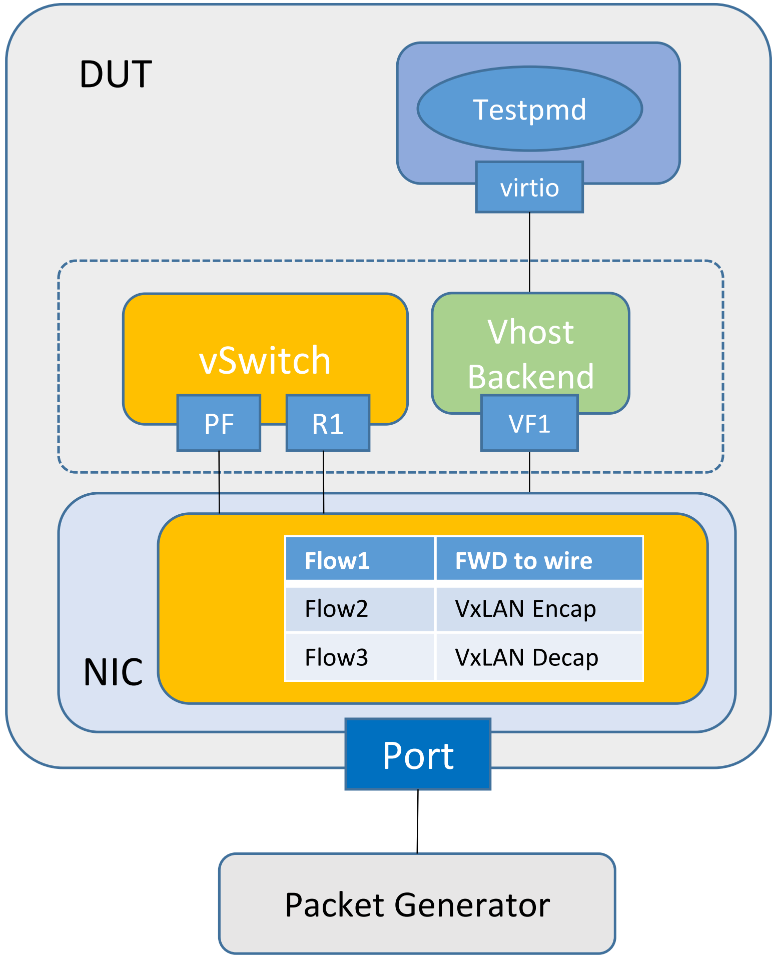
System Configuration:

(1) E5-2650 V4 @ 2.20GHz

(2) Mellanox 100G ConnectX-5 NIC

(3) Performance Metric: Mpps at 64byte

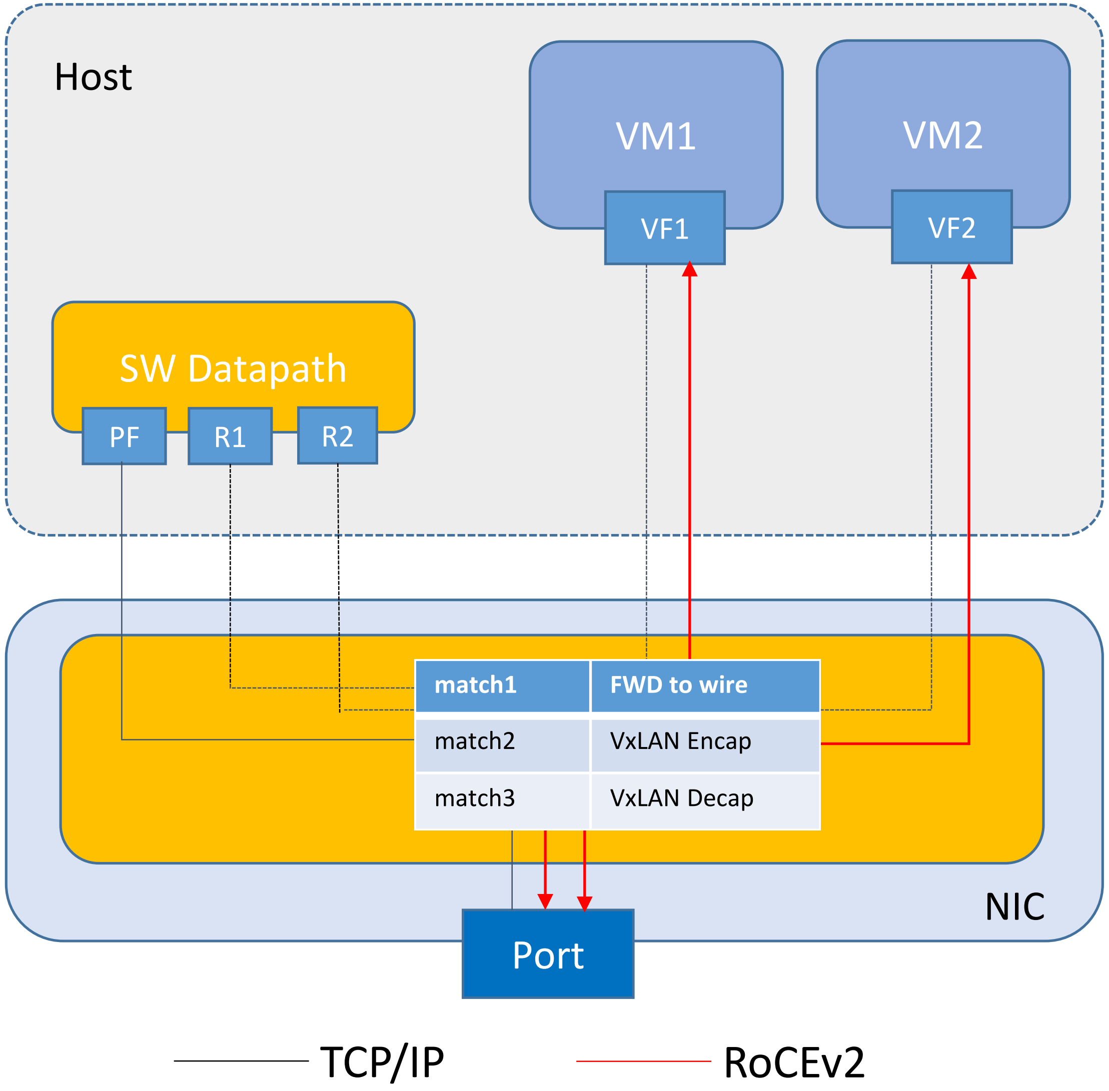
- Overhead caused by HW Encap/Decap on the path from/to VM can be minimized.
- If vhost-backend is strong enough, HW Offloading can bring high performance to virtio also.



Flow1	FWD to wire
Flow2	VxLAN Encap
Flow3	VxLAN Decap

RoCEv2 Support

- Only support SRIOV interface
- All RoCEv2 must go through HW datapath.
- ARP should be handled by SW datapath so that two endpoints of RoCEv2 can exchange address information.
- All RoCEv2 should be sent to wire or local VFs directly through configured rules like following
  - match {ip_proto=udp, dport=RoCE, dmac=<mac of VF1>} action {fwd to VF1}
  - match {ip_proto=udp, dport=RoCE, dmac=<mac of VF2>} action {fwd to VF 2}
  - match {ip_proto=udp, dport=RoCE} action {fwd to wire}
- New HW support VxLAN Encap/Decap for RoCEv2
  - Encap header can be based on inner (src ip, dst ip) + VNI.
  - ECN information need be copied from outer header to inner header after decap on RX.



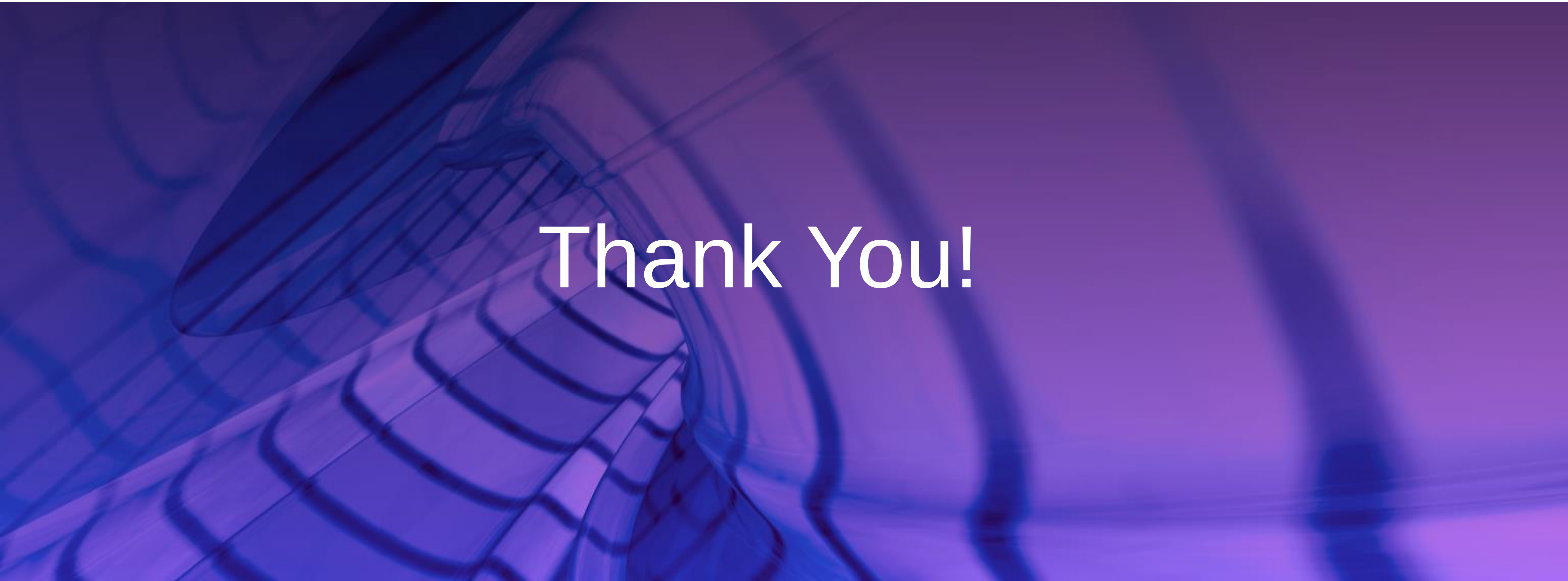
match1	FWD to wire
match2	VxLAN Encap
match3	VxLAN Decap

Other key Consideration

- **VF LAG**: VM sees only one VF while it can use two physical ports for Load balancing and link redundancy.
- **VF Mirroring**: mirroring the traffic from/to one VF to a dedicate admin VF for monitoring and traffic analysis.
- **Connection Track**: sending TCP packets with specific flags to software for processing connection state.
- **Live Upgrade**: update to new instance, need migrate both SW & HW datapath and interfaces from current instance to new instance.
- **SRIOV Live Migration**: VM with SRIOV VF can be migrated from one machine to another machine.

DPDK

DATA PLANE DEVELOPMENT KIT



Thank You!