IPsec Full Offload

Boris Pismenny

November 2017

Mellanox TECHNOLOGIES

Connect. Accelerate. Outperform.™

Overview

- Transparent IPsec is when HW provides a full IPsec data-path implementation:
  - ESP crypto, encap/decap, replay protection, sequence number generation, counters, notifications.

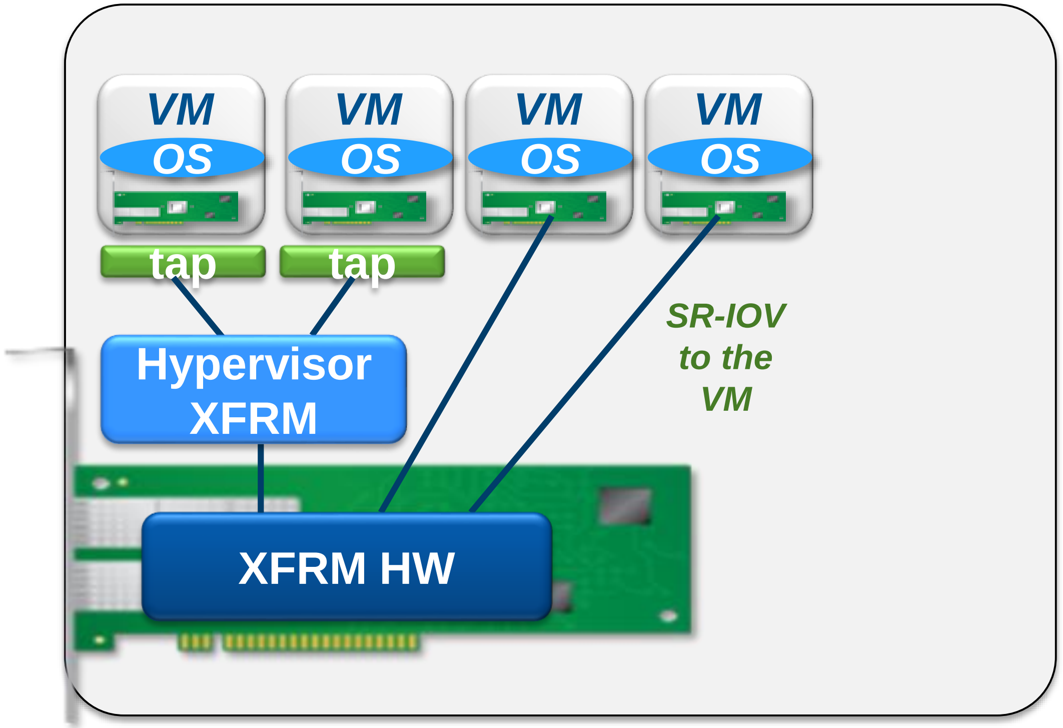
- There are two major use-cases:
  - Virtualization
  - Native Host

Full IPsec Offload - Virtualization

Virtualization

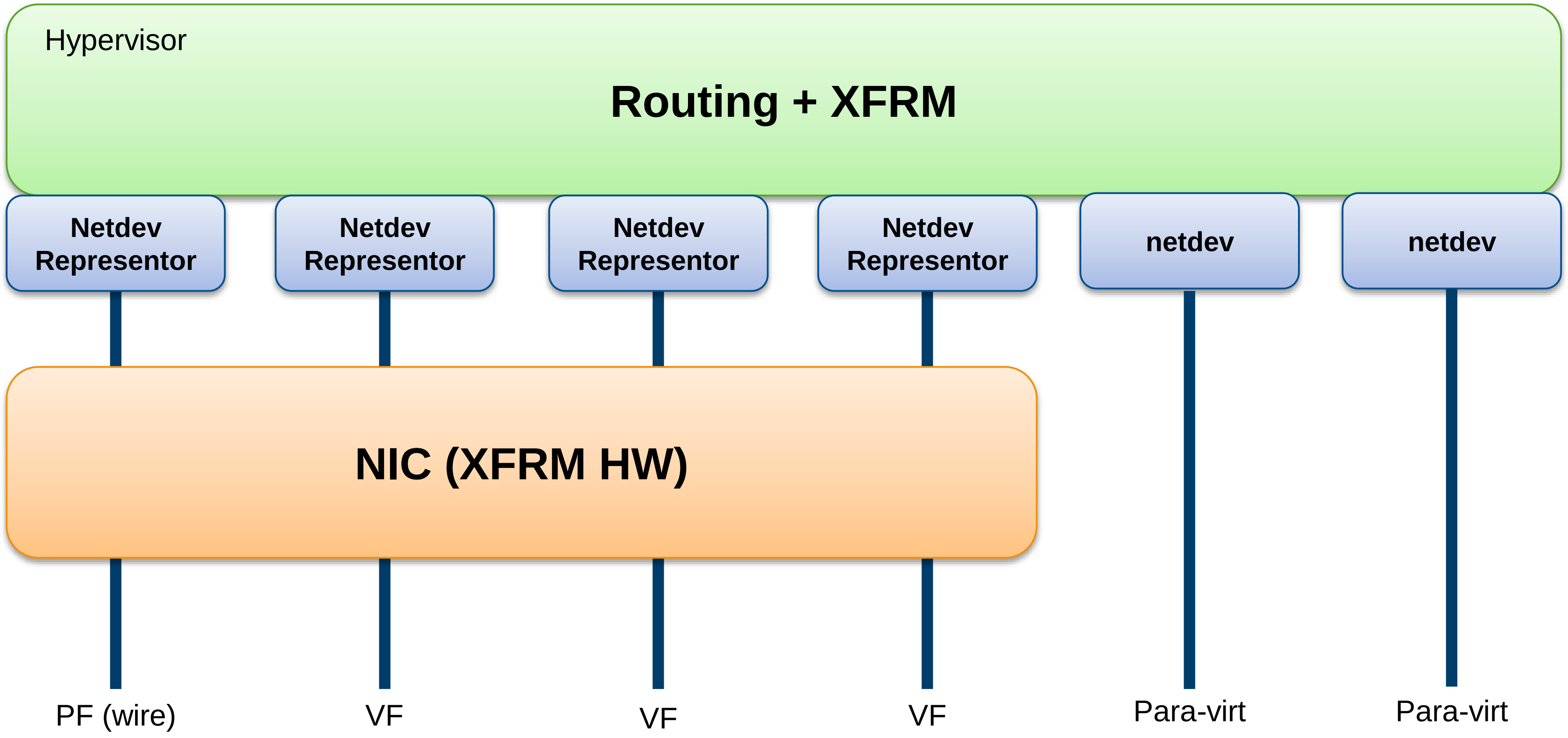
- The hypervisor xfrm layer is used to provide IPsec transparently for paravirtualized VMs

- Full offload could provide transparent IPsec for SRIOV VMs
  - The hypervisor (or Dom0) configures transparent IPsec for VMs



In-Host Network Topology

- Arch covers both cases of directly attached (VF) and Paravirt (PV) VMs
- To support the design we use VF representors
- Representor ports are a netdev modeling of eSwitch ports
- The VF representor netdev supports the following operations
  - Set IPsec Policy
  - Set IPsec SA
- The hypervisor could provide a software-fallback using the VF representor
- Packet sent on VF representor -> Packet received on VM VF NIC
- Packet sent on VM VF NIC -> Packet received on VF representor (except offload)



User Interface

XFRM netlink:

- Add the full offload option for XFRM_MSG_NEWPOLICY
  - Calls the netdev for new policy (works with representors too)
  - For representors:
    - Called before VM is created
    - Guarantee that the ACQUIRE packet will reach the XFRM stack on the hypervisor

- Add the full offload option for XFRM_MSG_UPDSA

Implications

- XFRM is used for the control plane and for software fallback
  - Most traffic bypasses the XFRM stack

- SA selector and policy checks are done in HW
  - Drop and count in HW on mismatch

- Soft/Hard limit events for Packets/Bytes must be generated by HW and propagated to XFRM
  - HW may not support both packet and byte limits
  - The hypervisor could provide the time limit, and check HW counters periodically to see if the SA is unused

- All statistics are provided by HW

- Auditable events will be provided to the hypervisor
  - Assisted by HW when necessary

- Limited configuration of replay protection, packet/byte limits
  - How to expose what is supported?

Exceptions

- Fragmentation (egress):
  - Transport mode SA needs to send an IP fragment – drop
    - Avoided by configuring tunnel mode SA when fragmentation is possible
  - Packet is bigger than MTU after adding all IPsec overhead – drop
    - Avoided by setting the MTU correctly

- Fragmentation (ingress):
  - All incoming IP fragments will be passed to the hypervisor's network stack for reassembly. Then handle in software fallback.

- Software fallback must update esn and replay protection atomically
  - Software could stop offload, dump HW state, and handle the packet

  or

  - HW could provide an atomic replay protection test_and_set (need to query ESN as well)

Note: It is possible to drop all exception packets in HW

Full IPsec Offload - Native

Native

- Not necessary to offload the policy check

Egress

- Packets must update the state in HW (even when rerouting or when using a bond)
- offload encap - skip most xfrm code
  - The network stack must see all headers, right?
  - Pass the headers through the network stack and remove them in the driver?

Ingress

- HW decapsulates ESP and checks replay
- Packets using a fully offloaded SA must come from the HW offload interface
  - Drop packets from other interfaces
- Driver would set the sec_path

© 2016 Mellanox Technologies

User Interface

XFRM netdevice:

- Add the full offload option for VTI/XFRMI

Limitations

- Packets **must** go through the offloading NIC in both egress and ingress
  - No bonding
  - No rerouting

- Received IP fragments are either dropped or trigger a software fallback

- Software fallback must update replay protection atomically (same as in virtualization)
  - Software could stop offload, dump HW state, and handle the packet

  or

  - HW could provide an atomic replay protection test_and_set

- Limited configuration of replay protection, packet/byte limits

Thank You

Mellanox TECHNOLOGIES

Connect. Accelerate. Outperform.™