


KubeCon | CloudNativeCon

Europe 2020

Virtual

Network Security for K8s Bare Metal Nodes

Girish Moodalbail, NVIDIA
Liel Shoshan, Mellanox (now NVIDIA)

Agenda

- Challenges Running Pods on Bare Metal
- Open Virtual Network (OVN) Primer
- OVN on SmartNIC (BlueField)
- OVN HW Offload
- SmartNIC Advantages

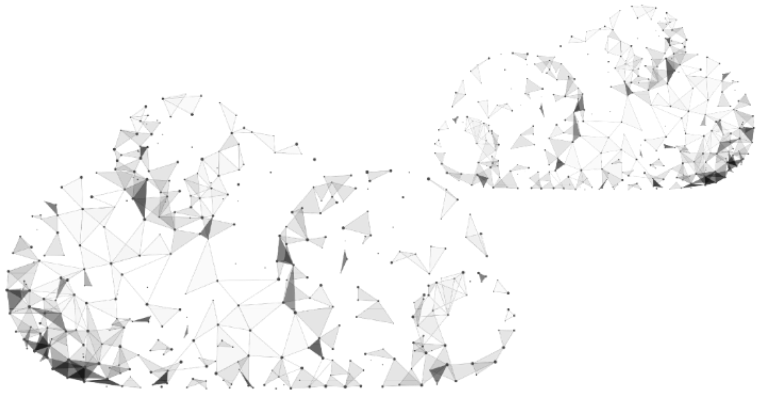
VM as Security Wrapper for Pods

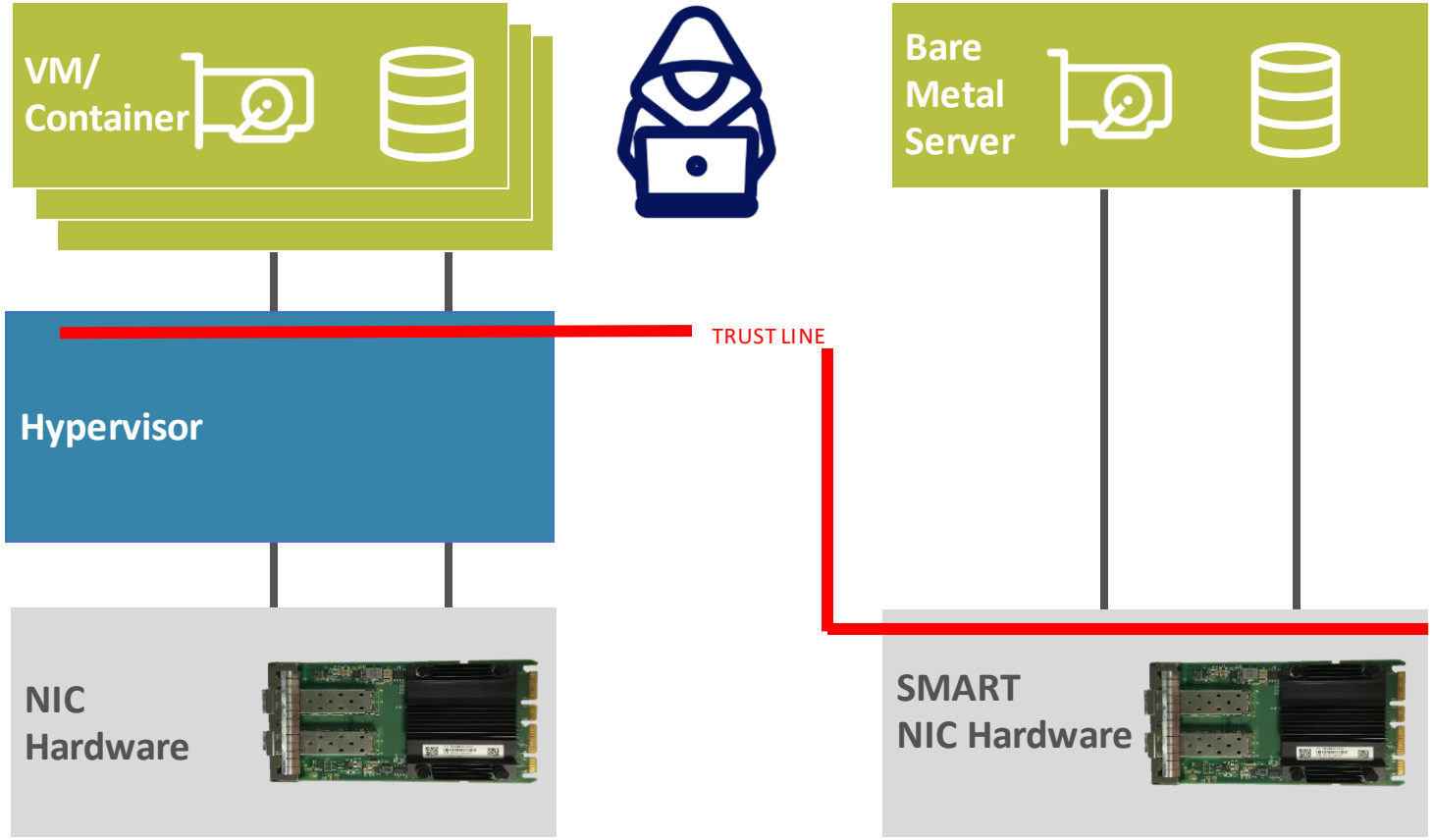
KubeCon CloudNativeCon Europe 2020 Virtual

- Kubernetes clusters are usually deployed on virtual machines
  - VMs are hosted either on public clouds or on-premise private clouds
- Operators are not confident about Container security
- If attackers broke out of the Container, they will still be trapped inside the VM
  - Therefore the data center will be safe
- Huge application performance cost:
  - Hardware resources need to be virtualized for VM access and then Container access
  - Hypervisor overhead for application performance

Bare Metal Platforms

- Driving Forces
  - Performance
  - Security and Isolation
- Trust zone shifts into the SmartNIC





Running Pods on Bare Metal

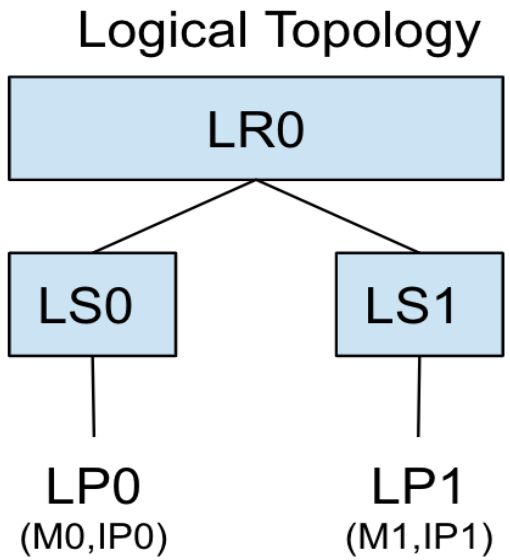
- Limiting blast radius of a compromised node through Pod escape is crucial for data center wide security
- Done through "bump in the wire" SmartNIC and distributed SDN control plane
  - Open Virtual Network (OVN) control plane, offloaded Open vSwitch (OVS) data plane
- The K8s APIs are implemented using OVN Logical Resources
  - Both standard APIs and extensions through CRD
- The datapath and security policies for the Pods are implemented as OVS flows
  - The OVS OpenFlow flows are accelerated on the Smart NICs

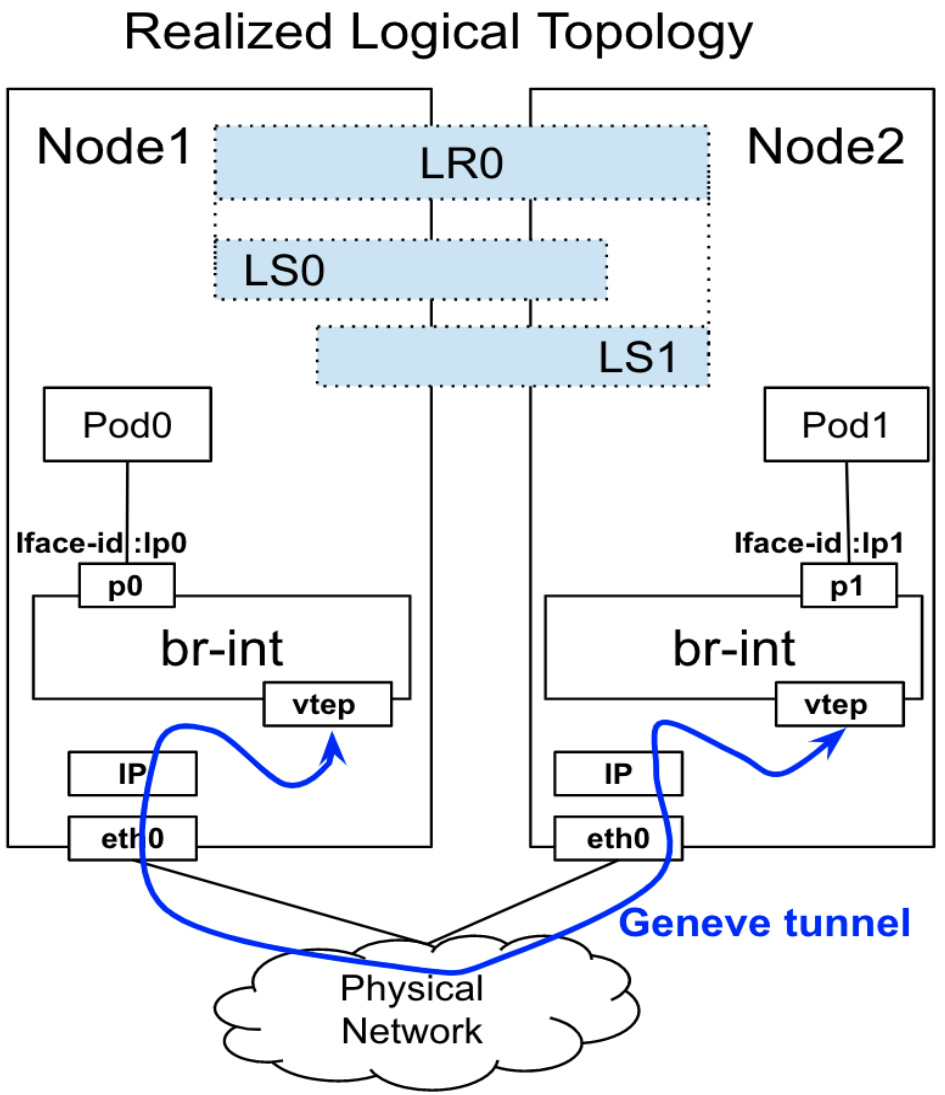
Open Virtual Network (OVN)

- OVN is an opensource network virtualization solution developed by the Open vSwitch community (https://github.com/ovn-org/ovn/)
- An Open vSwitch based solution
  - L2/L3 virtual networking
  - Logical switches and routers
  - Stateful Network policies
  - GENEVE overlay
- Mellanox NICs provides full datapath HW offload for OVN solution
  - Reducing CPU utilization from ~100% to ~0%
  - Increasing throughput drastically

OVN Logical Topology

LR → Logical Router
LS → Logical Switch
LP → Logical Port
iface-id: lp0 → Physical binding of a logical port to a node
VTEP → virtual tunnel endpoint

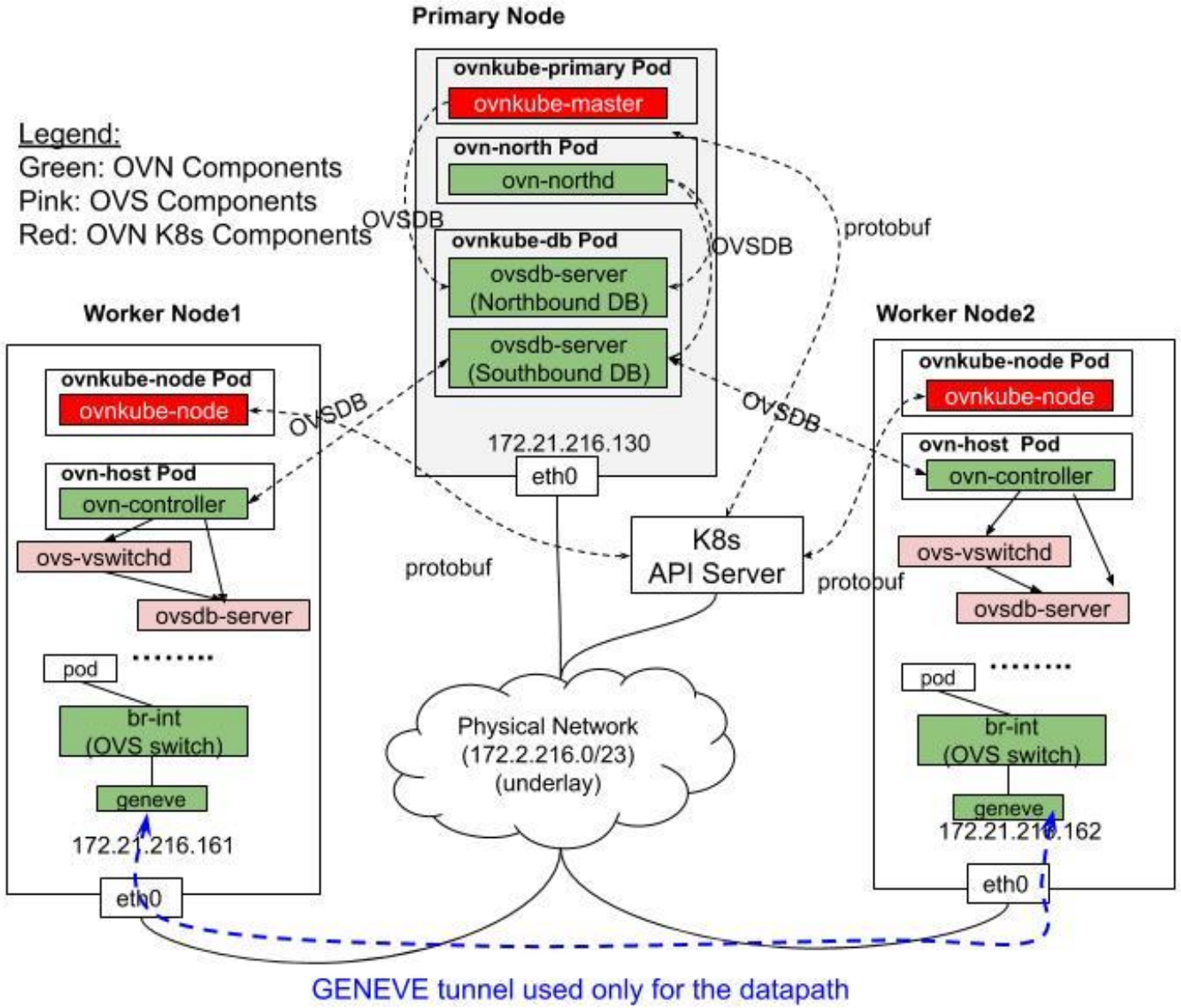




OVN Kubernetes Control Plane

The OVN project has a specific Kubernetes network plugin called ovn-kubernetes (https://github.com/ovn-org/ovn-kubernetes/)

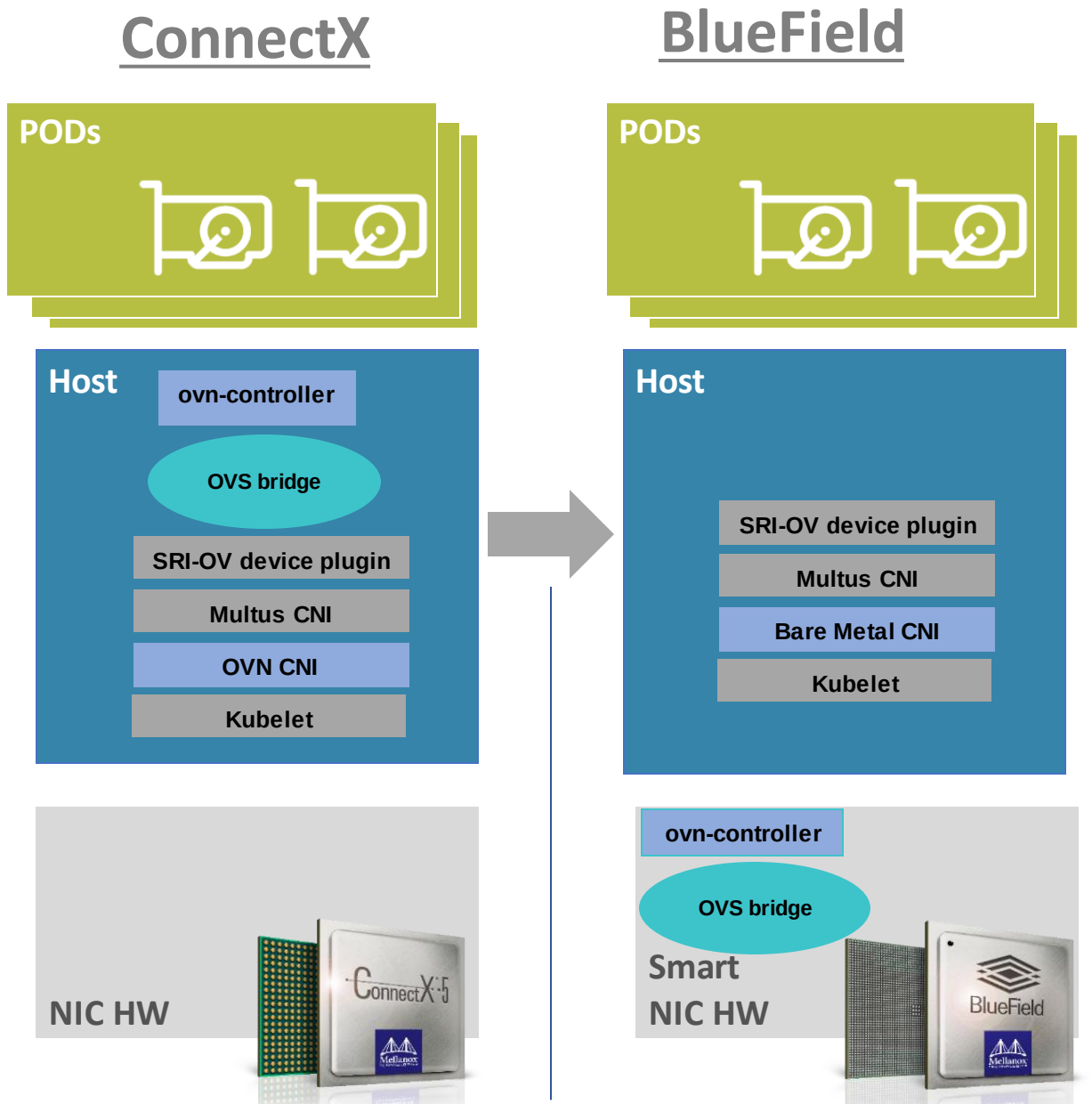
Layer	Architecture	Components
OVN K8s CNI	Client/Server (K8s Watcher)	ovnkube-primary Pod ovnkube-node Pod
OVN	Client/Server	ovn-db Pods ovn-north Pod ovn-host Pod
OVS	Standalone	ovs-vswitchd ovsdb-server



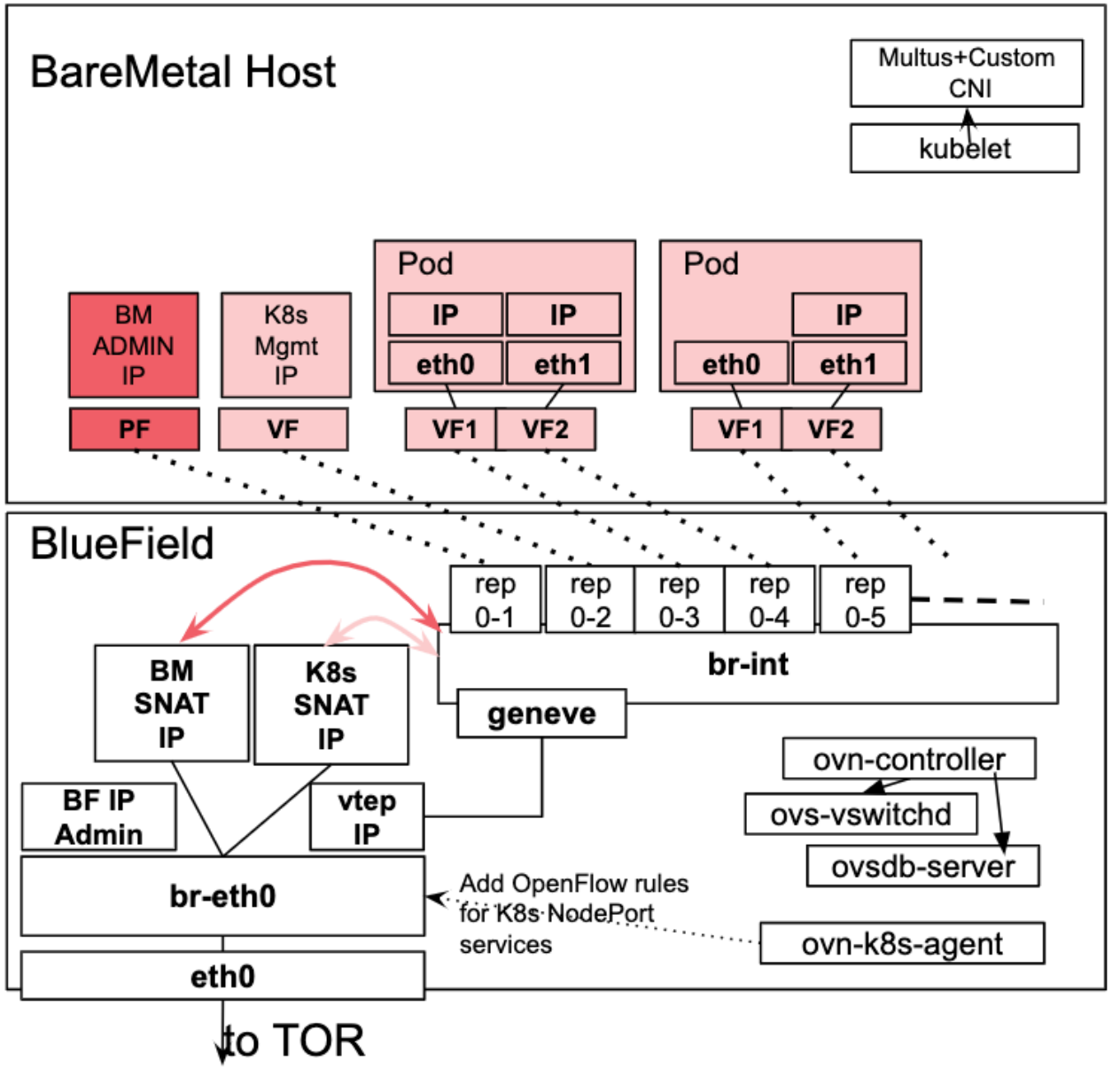
OVN K8s with Mellanox SmartNIC

Secured solution using BlueField SmartNIC:

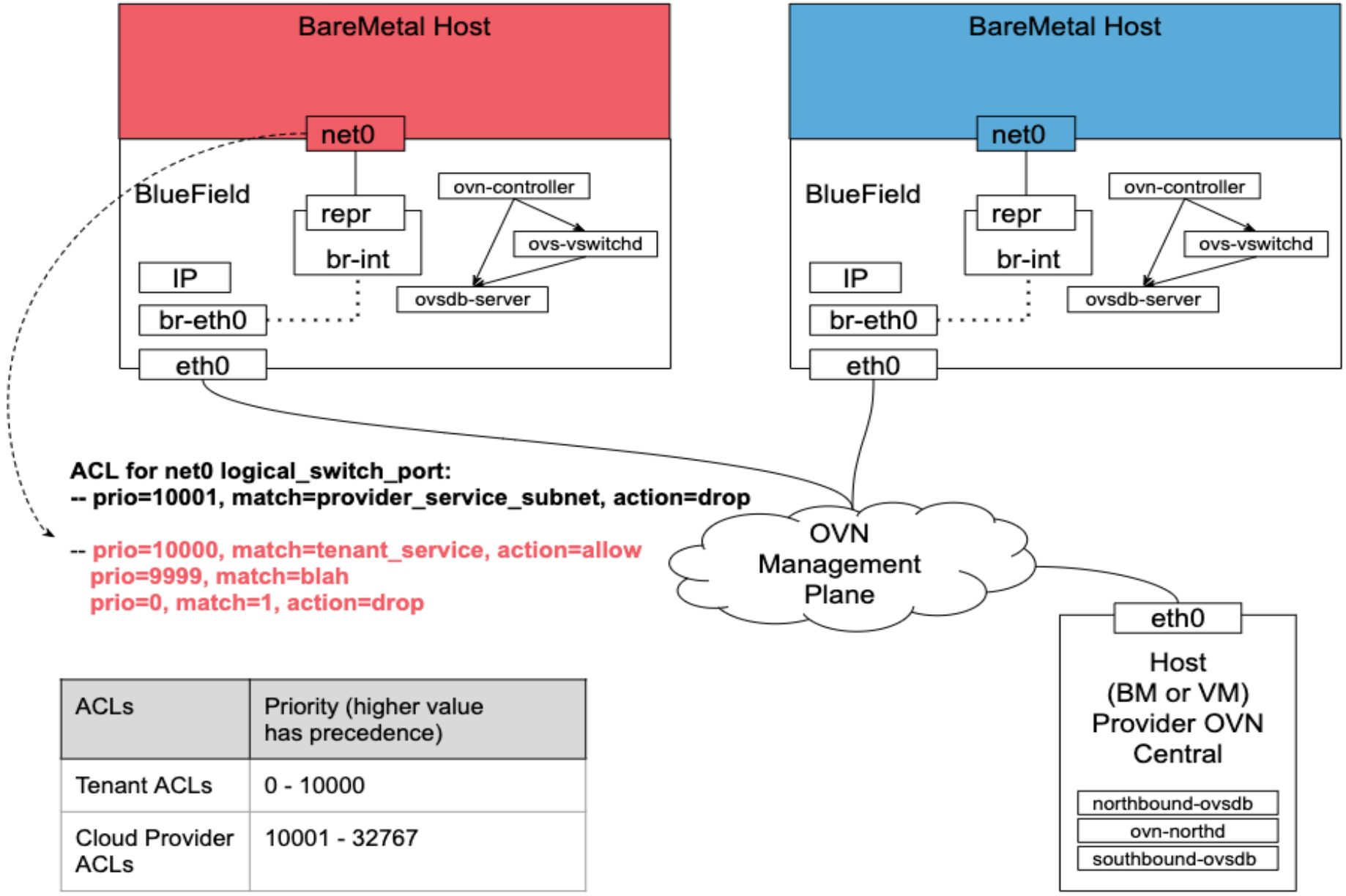
- OVS switch and the utilities configuring it will run on the ARM cores
- Increased performance and isolation



OVN Control Plane on SmartNIC

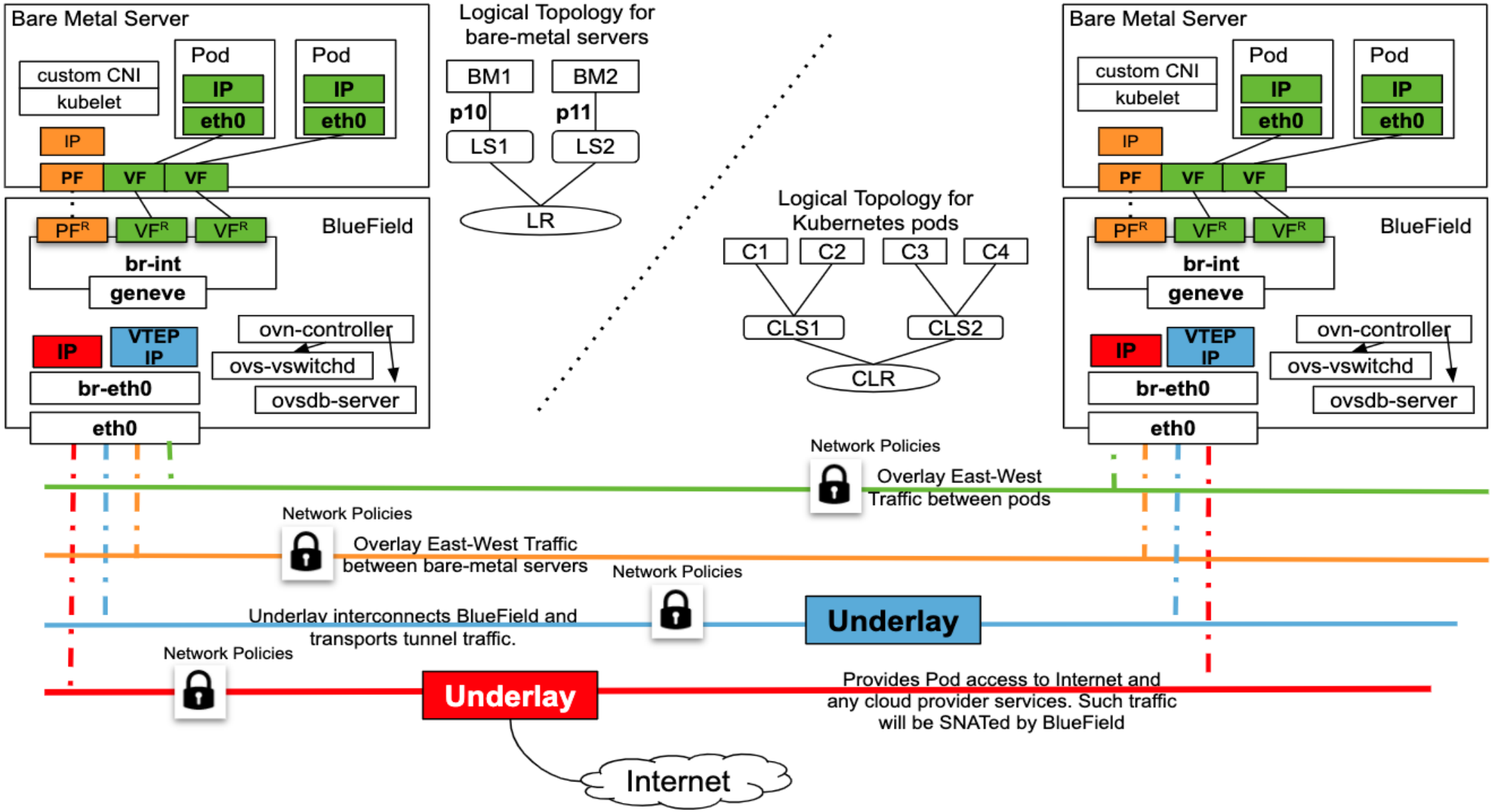


OVN Network ACLs

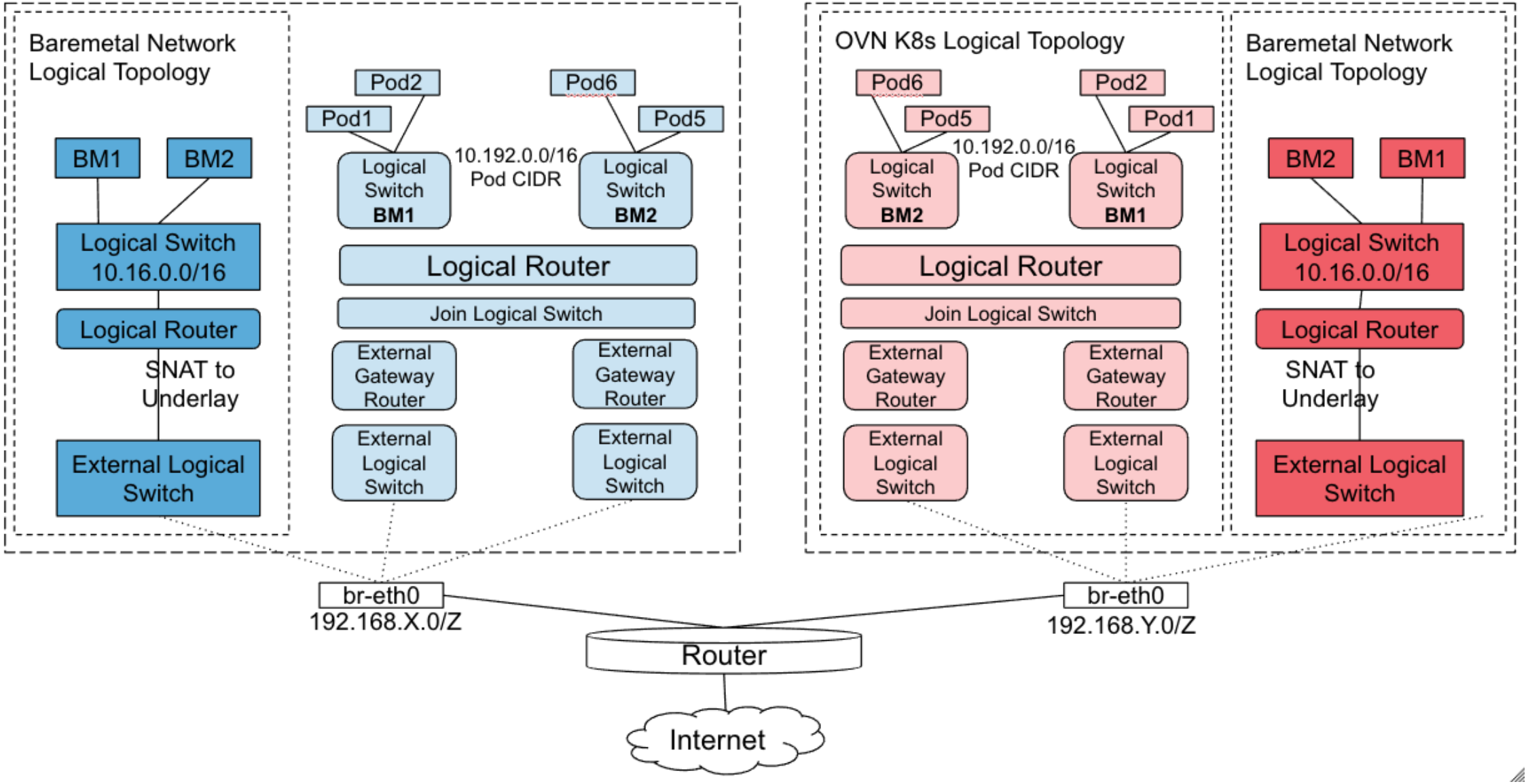


ACLs	Priority (higher value has precedence)
Tenant ACLs	0 - 10000
Cloud Provider ACLs	10001 - 32767

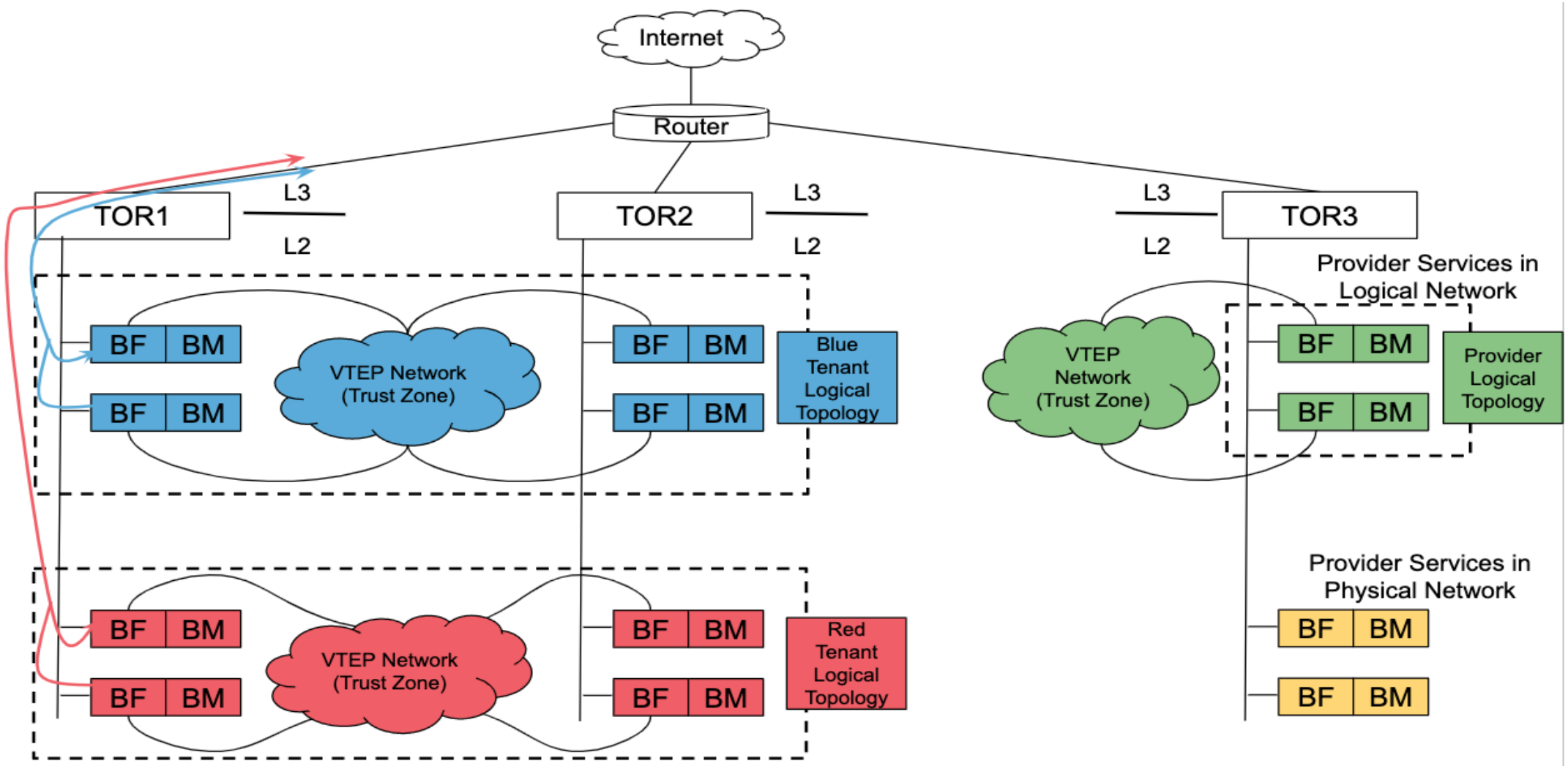
DC and Host Networking



Multi-Tenancy OVN Logical Topology



Multi-Tenancy DC Networking



Note: Transport (aka trust zones) are created amongst the tenant's chassis by running the following on each of the chassis

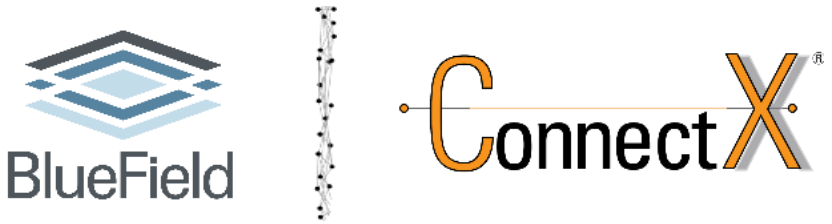
$ ovs-vsctl set open . external-ids:ovn-transport-zones=red

Introducing: ASAP²

- Accelerated Switching & Packet Processing
- A SW and HW integrated solution which utilizes Mellanox SmartNIC to accelerate and offload the Data-Plane
- Maintains control plane in Software
  - Minimize K8s CNI or SDN changes
- Support different customer configuration
  - SR-IOV or VirtIO
  - Control plane running in Kernel or in User Space (DPDK)
  - Accelerates customers' custom Virtual Switches/Routers or known open source solution (OVS, Tungsten Fabric, etc.)
- Upstream and Inbox solutions

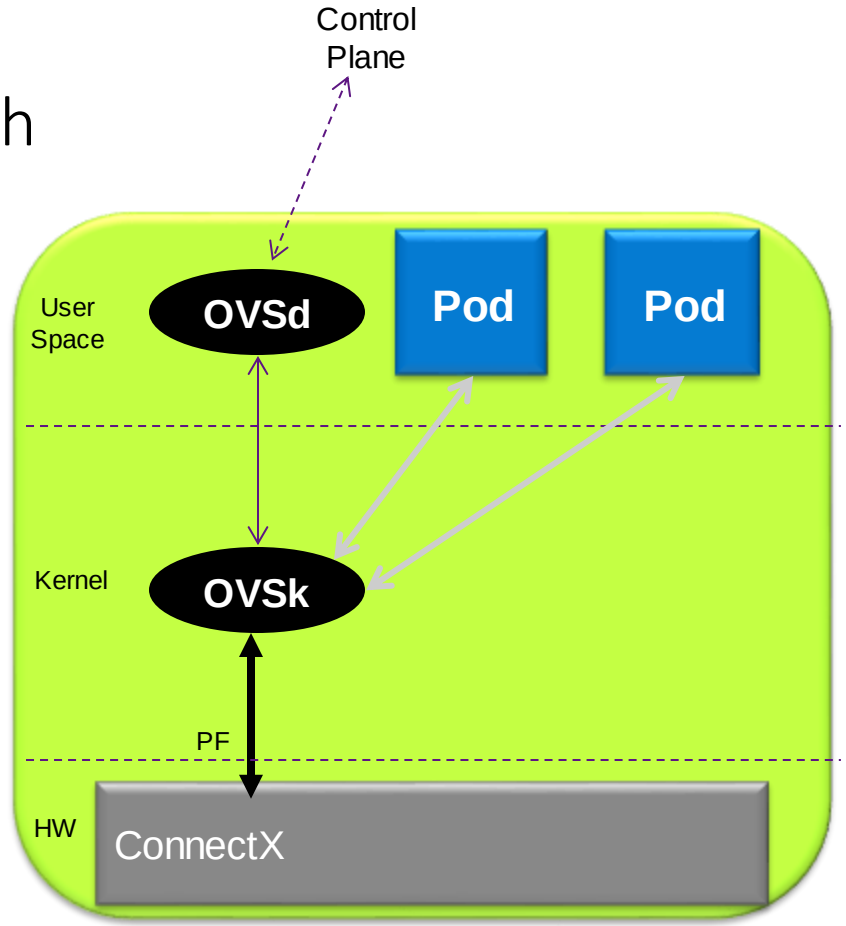
BlueField | ConnectX®

- Hardware offloads to virtual switches
  - Classification offload
  - Action offload
  - Datapath offload



OVS Offload With ASAP² SR-IOV

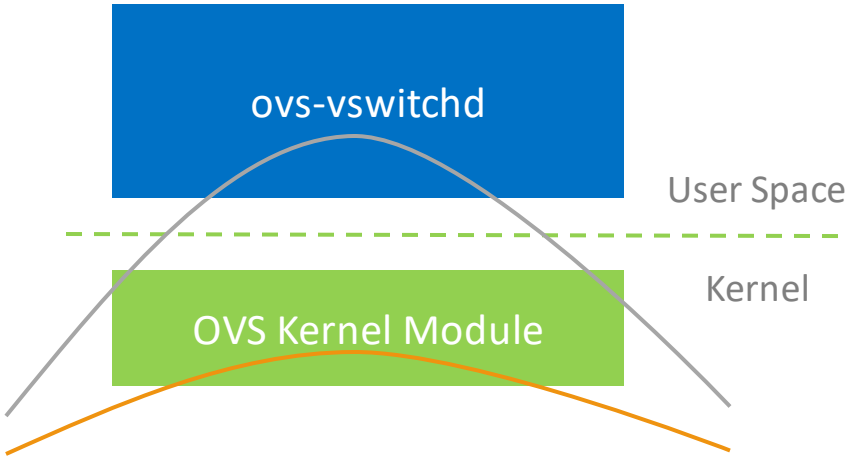
- Open vSwitch (OVS) is the most popular virtual switch
- Flow based switch
- L2, L3, NAT, VLAN, VxLAN, Mirroring, CT, and more
- Multiple control planes are available



Hardware based Acceleration

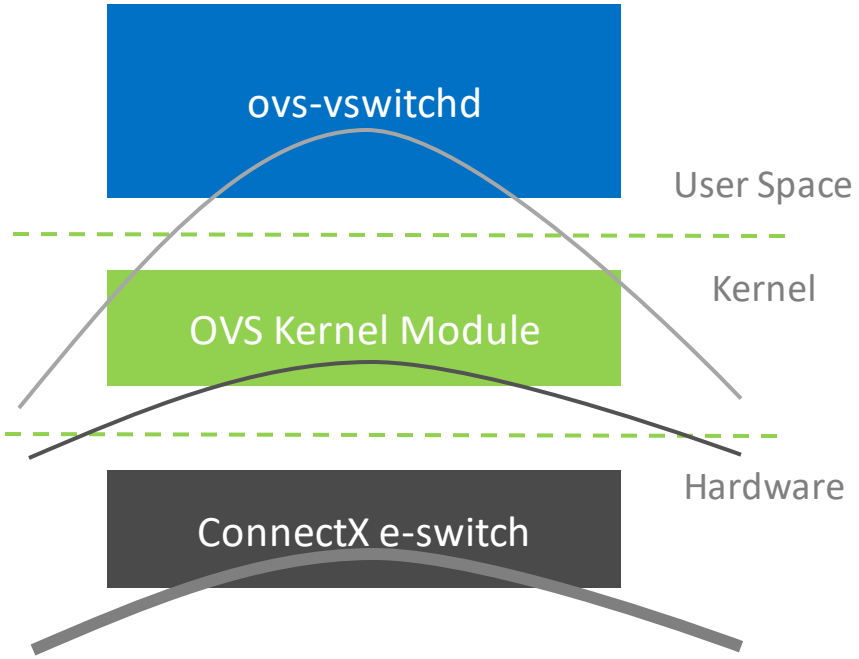
Traditional Model: All Software*

High Latency, Low Bandwidth, CPU Intensive



ConnectX: Hardware Offload

Low Latency, High Bandwidth, Efficient CPU



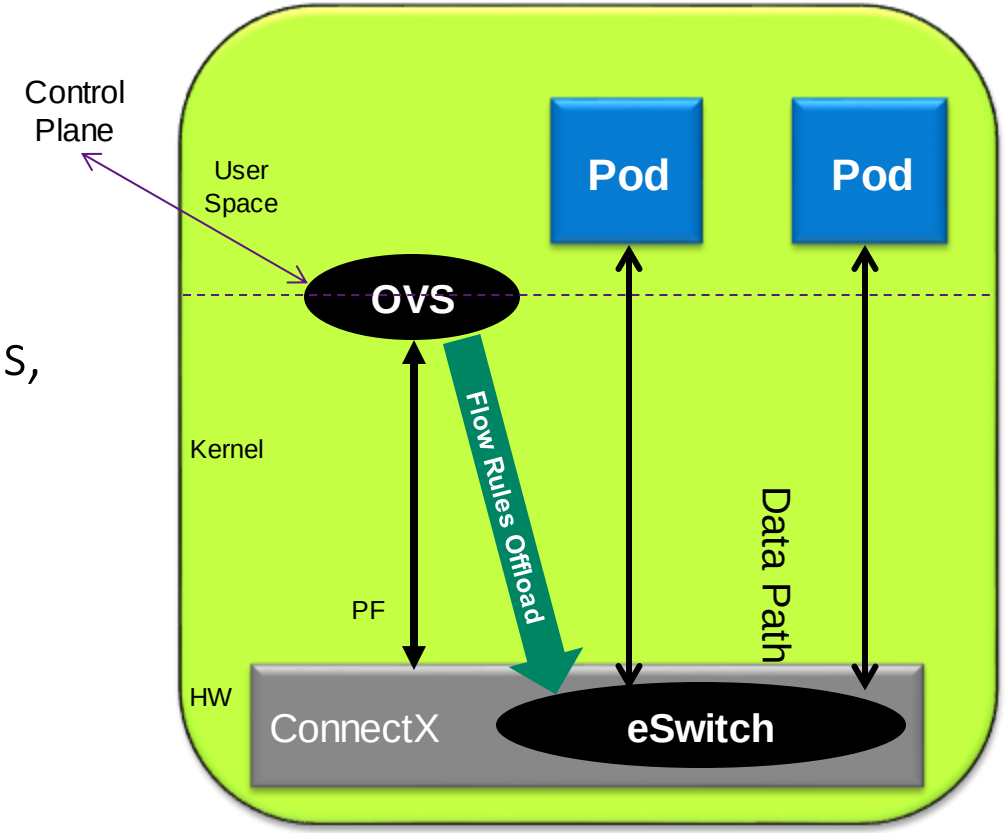
First flow packet

Fallback FRWD path

HW forwarded Packets

OVS Offload with ASAP² SR-IOV

- Keeps “first packet miss” arch with additional HW layer
- Packet delivery via SR-IOV
- OVS set the policies, eSwitch execute
- Linux TC rules, originating from OVS OpenFlow rules, are offloaded to the HW



Network Virtualization Using NIC

ASAP² Benefits

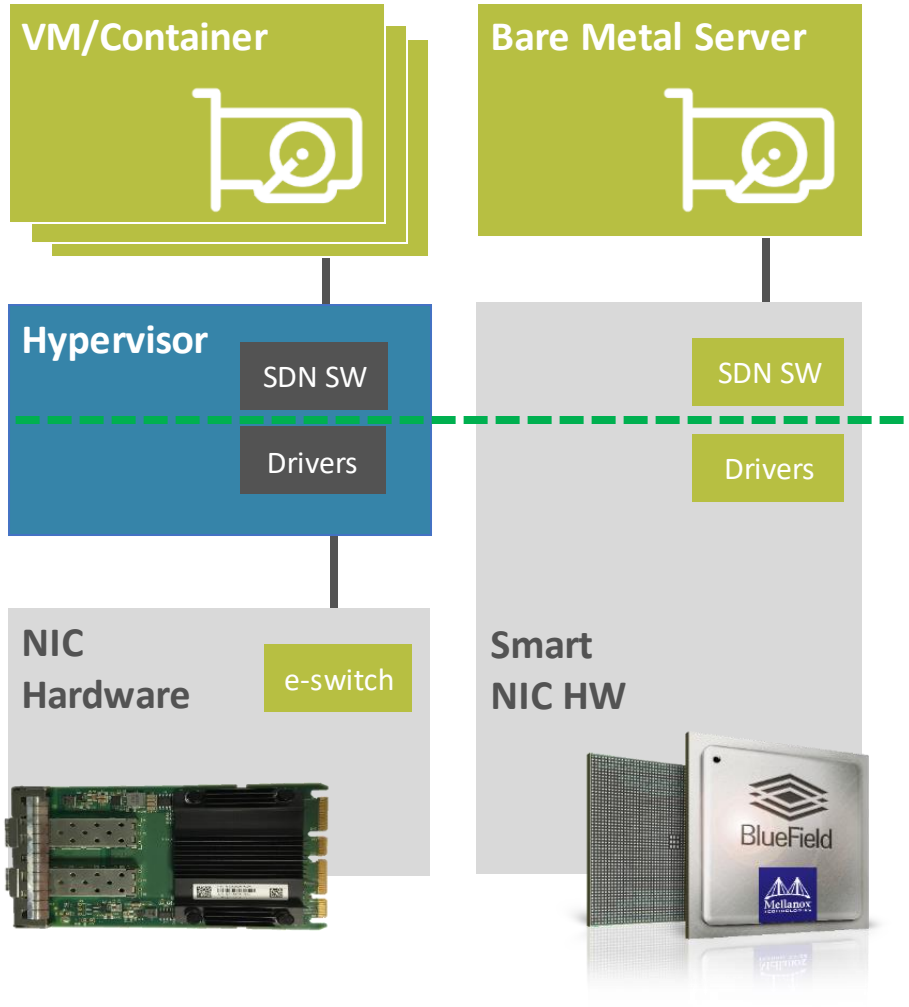
- Uncompromised performance
- CPU savings
- Full isolation
- Same solution for VM and BM
- OS/HV agnostic
- Security extensions

Opensource standard

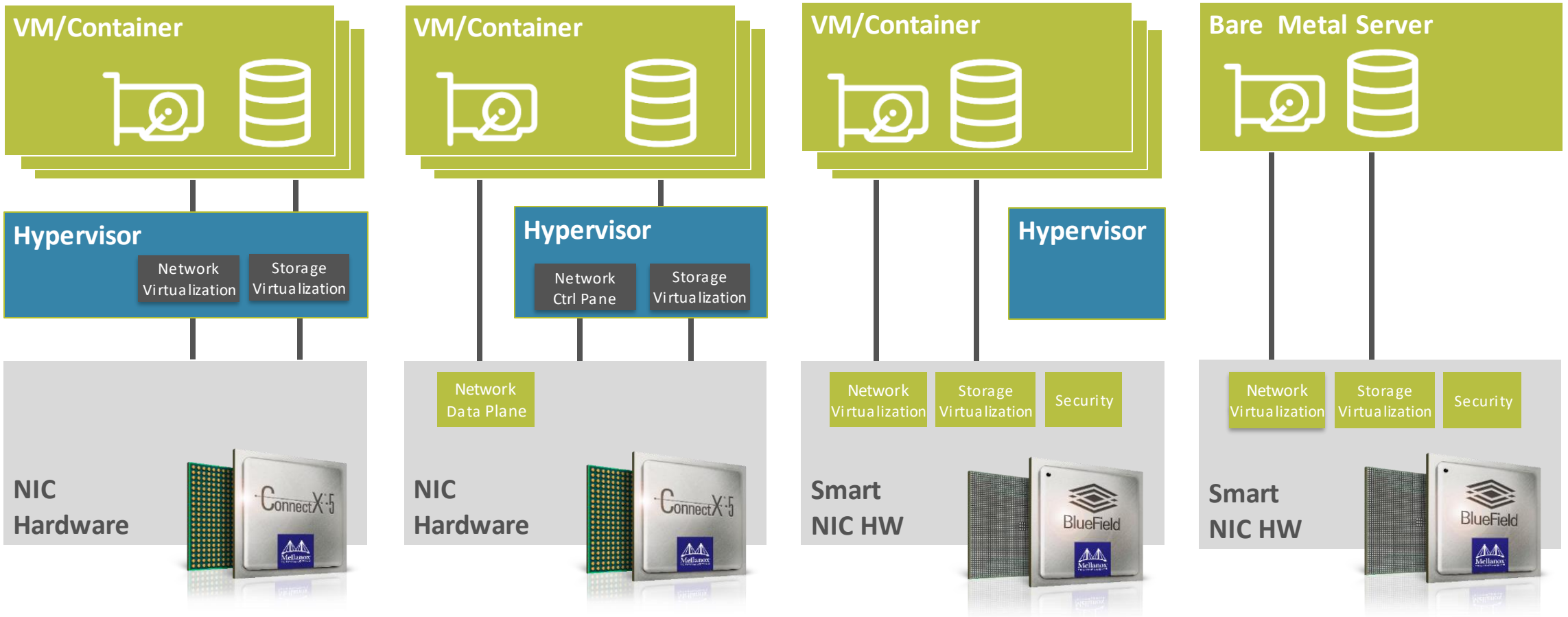
- Linux Kernel TC
- DPDK rte_flow

NETWORKING SECURITY

ASAP²
Accelerated Switching & Packet Processing



Software Defined Network, Storage, Security Transition




KubeCon
CloudNativeCon
Europe 2020
Virtual
HELM
KEEP CLOUD NATIVE
CONNECTED