Implementing virtual network offloading using open source tools on BlueField-2

“Simplified Packets”

Rashid Khan
Director Networking
Red Hat

Rony Efraim
Staff Architect
Nvidia

Red Hat

Who

Rony Efraim is Staff software architect at NVIDIA / Mellanox.

He is in charge of defining software & system architecture of NIC Acceleration for Virtualization. with emphasis on Ethernet, InfiniBand ,RoCE, SR-IOV, eSwitch, Open Vswitch (OVS), SDN, virtualization, OpenStack, K8s, DPDK, VNF, and Linux Kernel..
He has been working with many open source projects like OpenVswitch, DPDK, Linux networking for HW acceleration for smart nics.
He has +20 years of experience in the networking industry particularly with very large Telcos and clouds in NA, EMEA and APAC.

Rashid Khan is Director of Networking at Red Hat.

He is in charge of Linux Kernel Networking, ebpf / xdp, DPDK, OpenvSwitch, OVN, Smartnics, and Openshift Networking.
He has been working with many HW vendors on open source smart nics solutions, on Telco 5G, and NFV.
He has ~24 years of experience in the networking industry particularly with very large Telcos in NA, EMEA and APAC.
He also has vast experience in Audio and Video compression and broadcasting for streaming, broadcasting, unicasting and video conferencing

Red Hat

What we'll discuss today

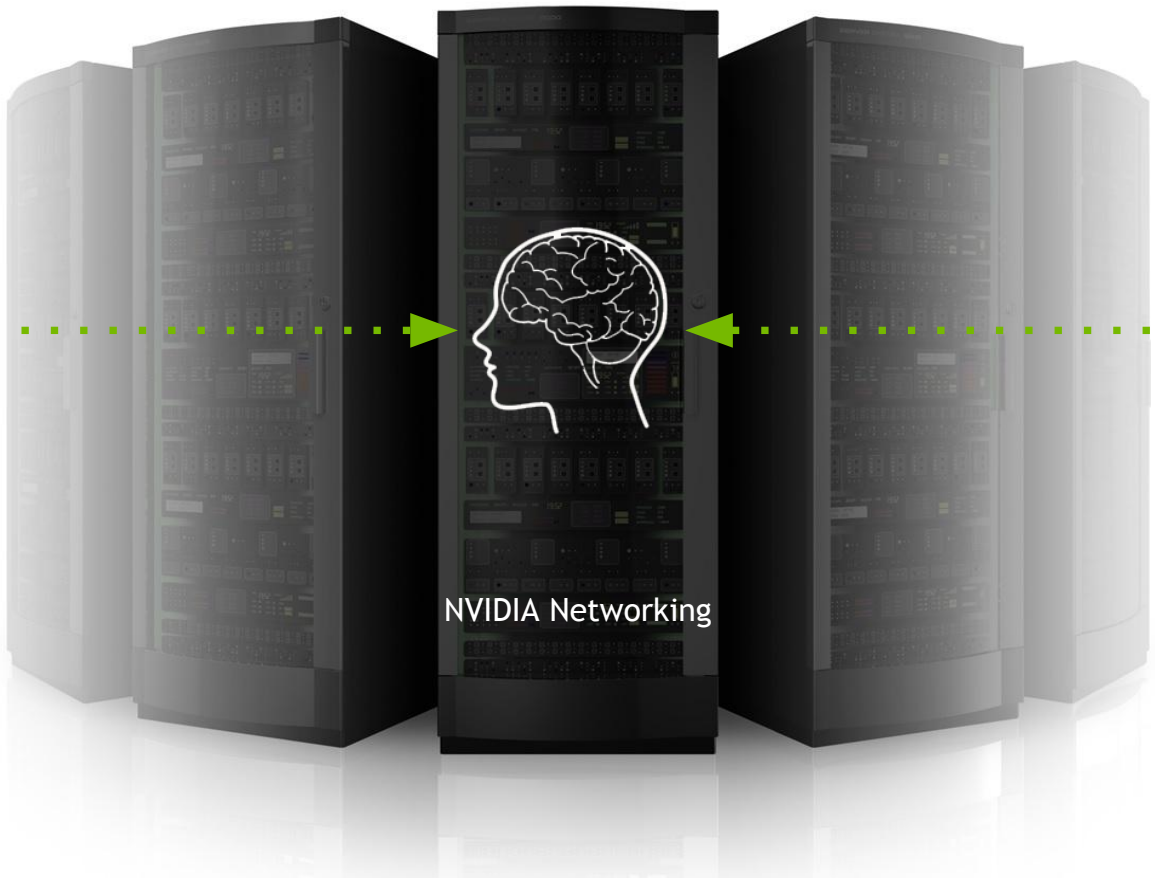
- Bluefield-2
- Geneve
- IPsec
- SR-IOV
- OpenShift
- Open vSwitch
- ovn-kubernetes

REINVENTING THE DATA CENTER

The Data Center is the New Unit of Computing

GPU

Accelerate Commuting
AI & Machine Learning

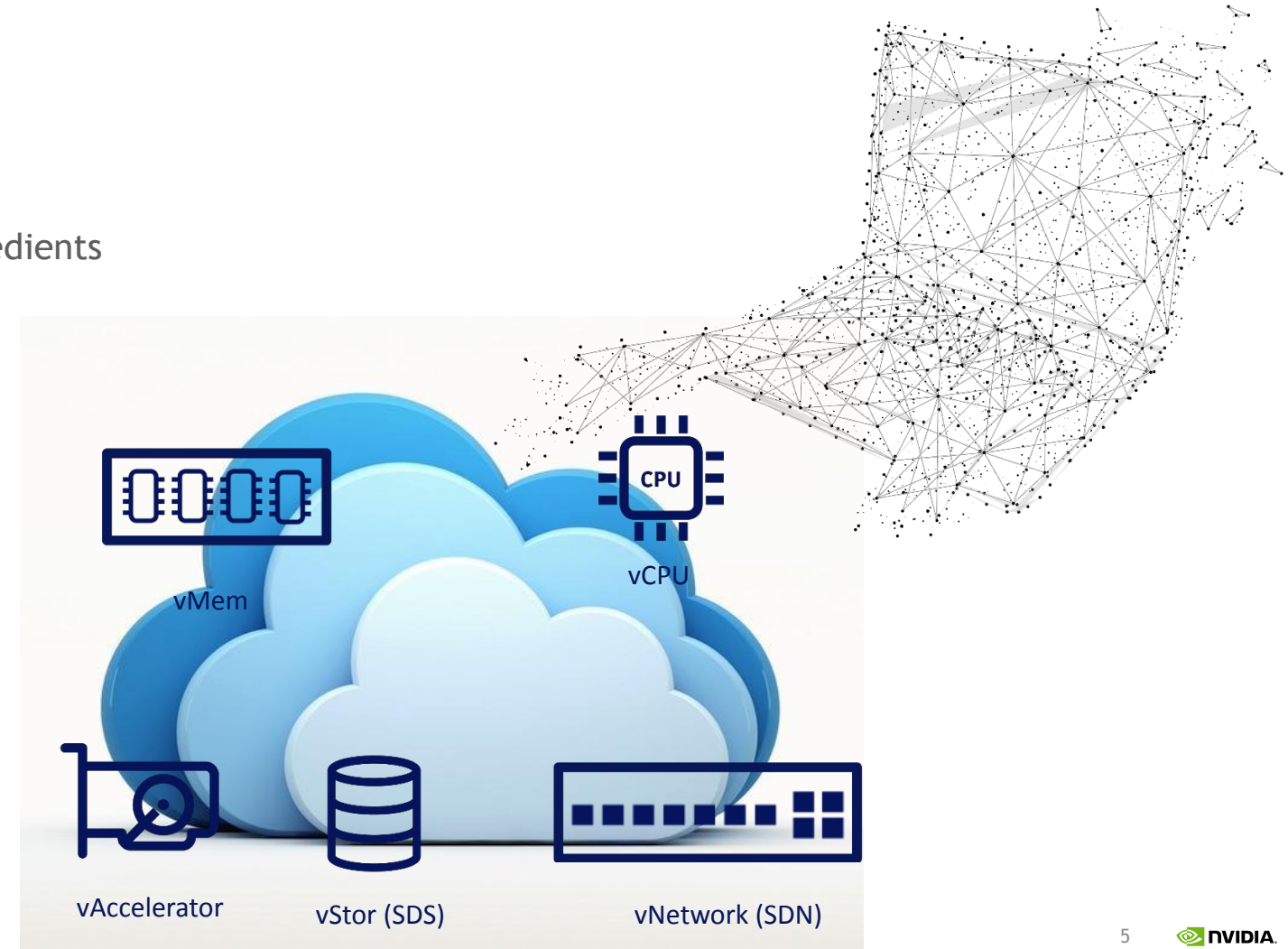


DPU

Software defined,
Hardware-accelerated

THE SOFTWARE DEFINED CLOUDED DATA CENTER

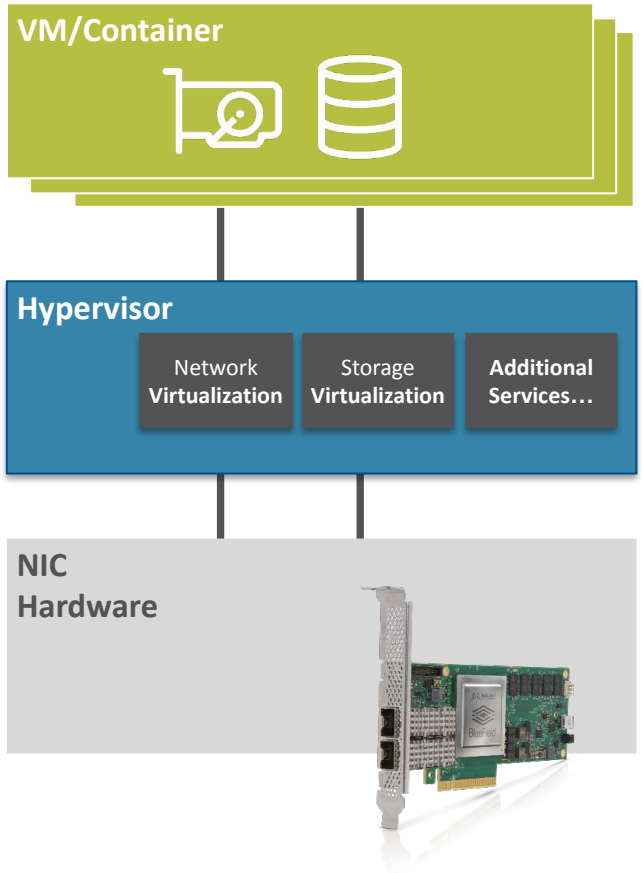
- Resource virtualization & disaggregation
  - Virtual instances comprise of physical ingredients
- Efficient services, Container & VM friendly
- Tenant isolation & security
- Visibility and telemetry
- Edge ready
- AI powered



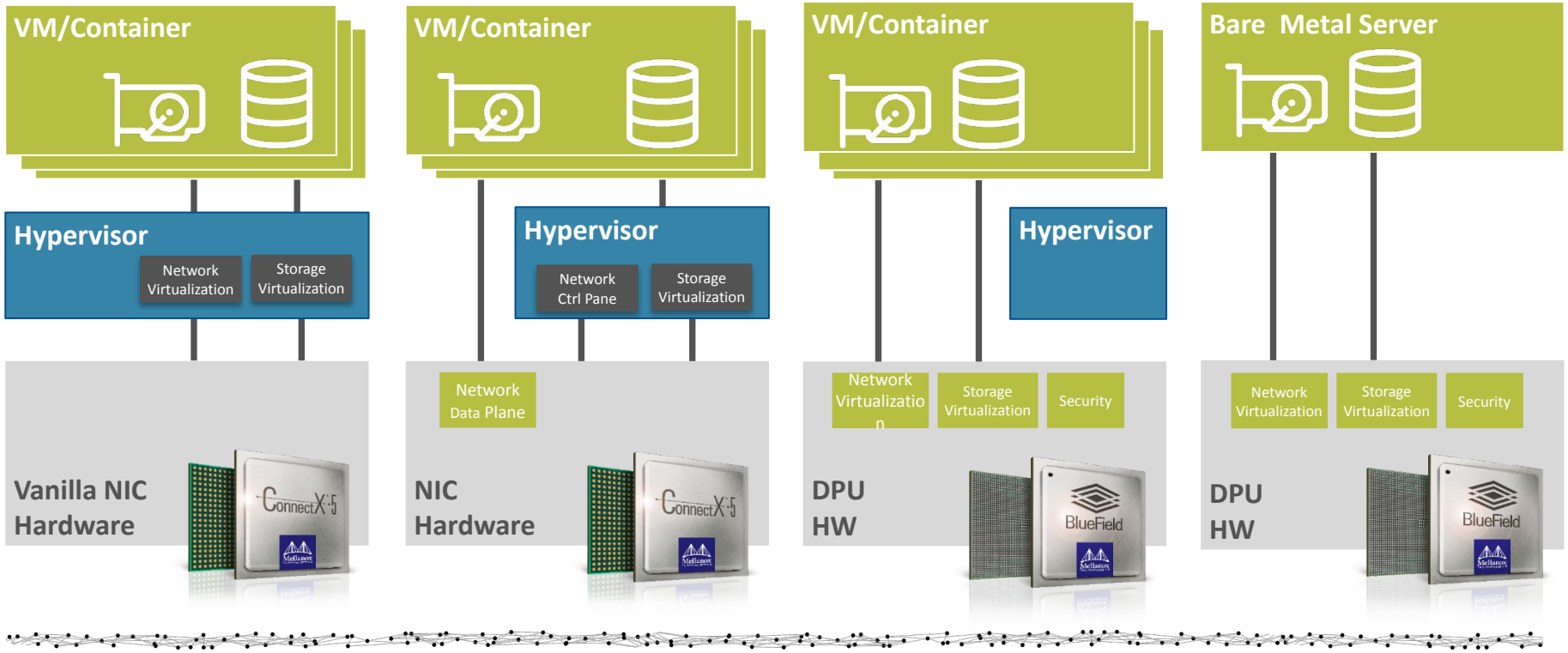
THE VIRTUALIZED DATA CENTER

Network & Storage Baseline Services

- Software Defined
- Scalable
- Secure
- Efficiency & Performance



SOFTWARE DEFINED NETWORK, STORAGE, SECURITY TRANSITION



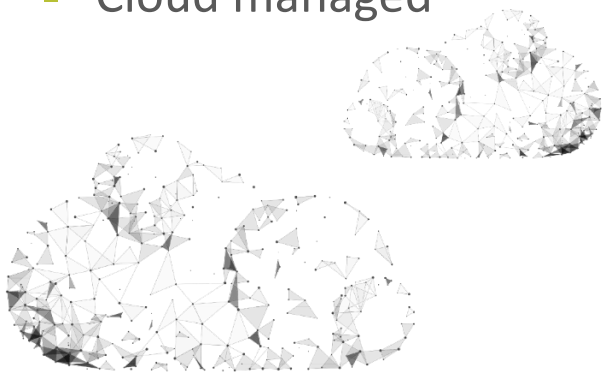
BARE METAL PLATFORMS EMERGENCE ▫ SMARTER NIC

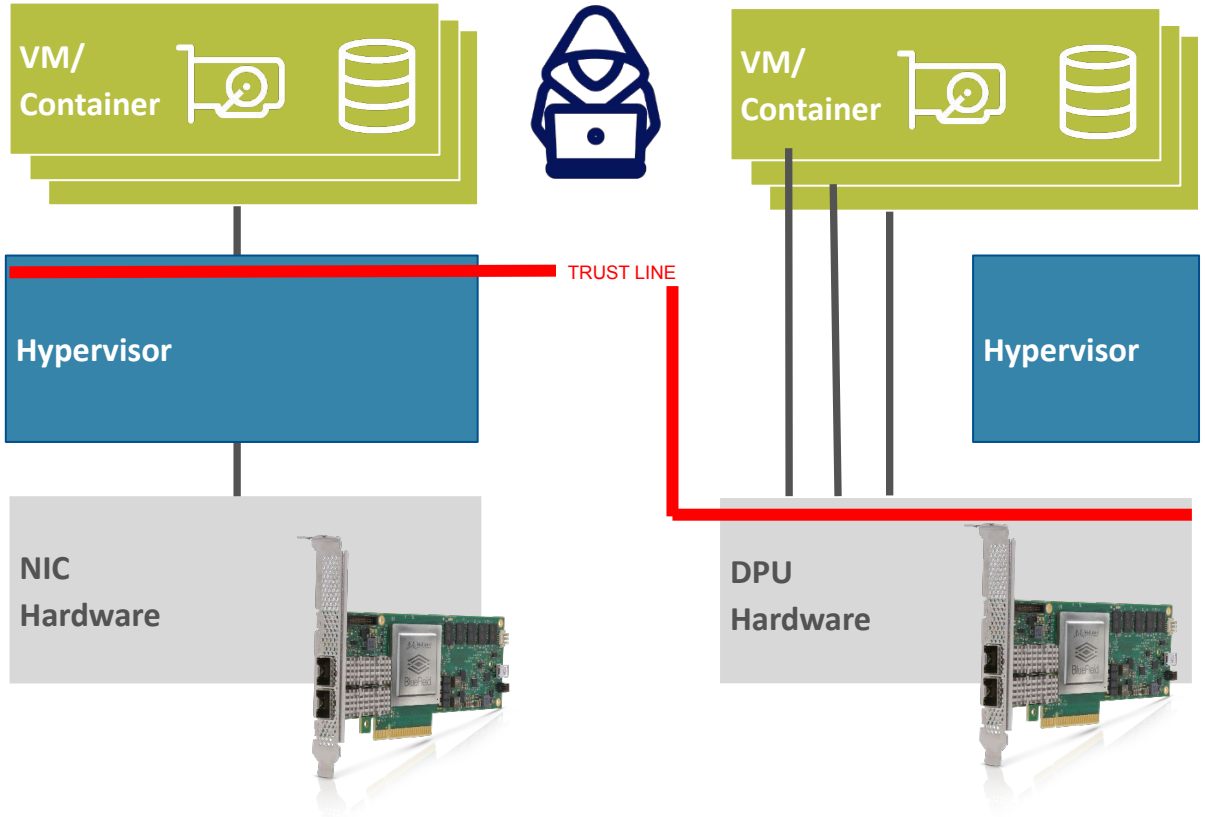
Driving Forces

- Performance
- Security and Isolation

Trust shifts into the DPU

- Cloud managed





DPU IN THE CLOUD

Fully managed CPU and DPU centric data center

- Provisioning – CPU and DPU day zero deployment and lifecycle
- Orchestration – Manage CPU workloads and DPU services
- Visibility – Telemetry and Data Analytics for CPU and DPU

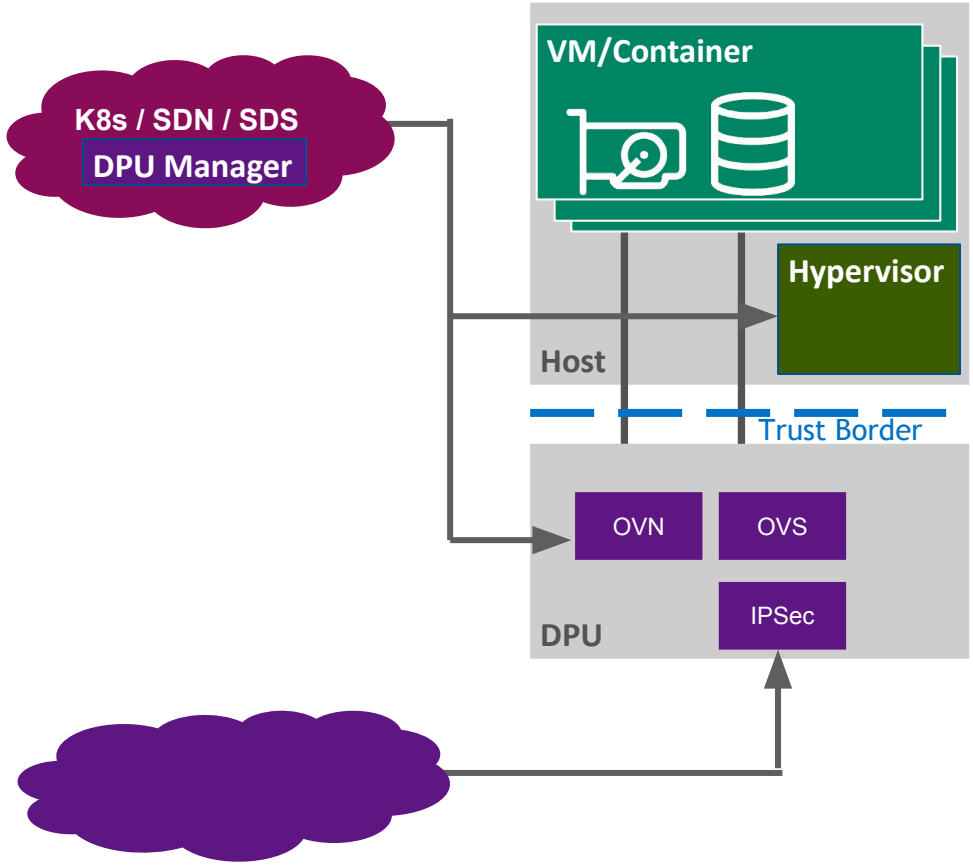
Cloud OS integrated with DPU management

Cloud, on-prem and hybrid. Multi tenancy

Network Accelerated apps and services for data center

Secure data center, SDN isolation

Integrated single controller for CPU and DPU



IPSEC INLINE ENCRYPTION OF DATA-IN-MOTION

Encryption/decryption at 100Gb/s bidirectional

- Lower CPU utilization with significant higher performance
- Protocol encapsulation and data plane accelerations (aware/un-aware modes)

Inline acceleration

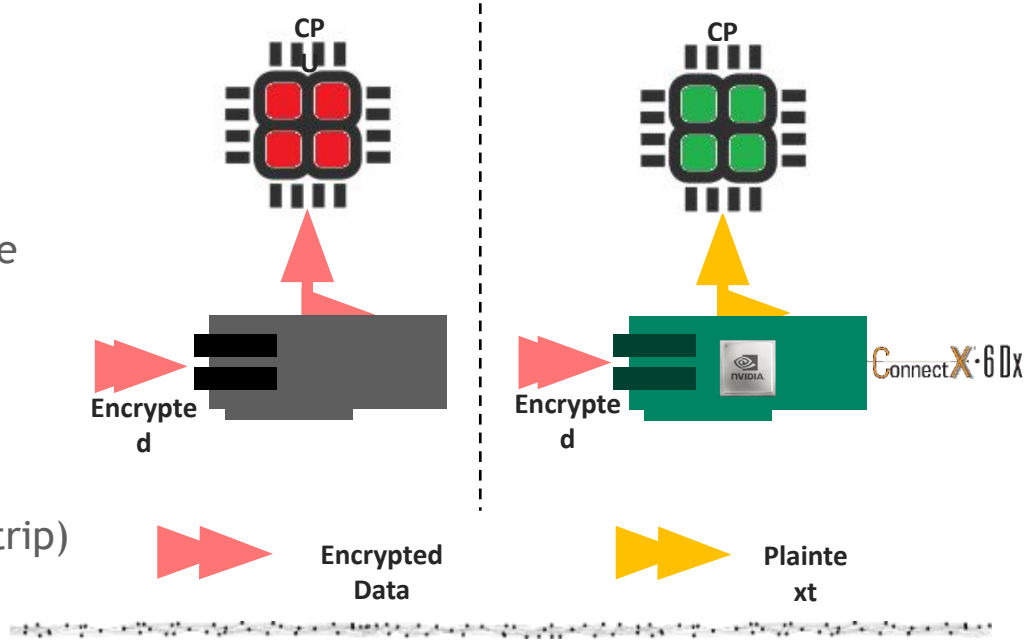
- Inline with other accelerations (tunneling, OVS, SR-IOV etc.)
- Removes software overhead of invoking accelerator (lookaside roundtrip)
- IPsec key management in software

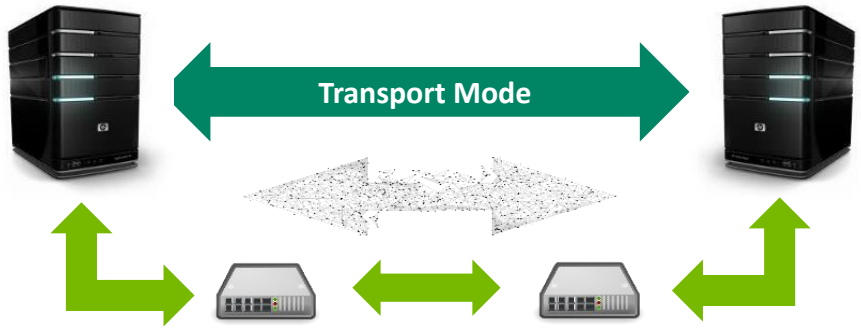
Support Transport mode and Tunnel mode

Use-cases

- Ethernet and RoCEv2 IPsec
- East-west data center encryption

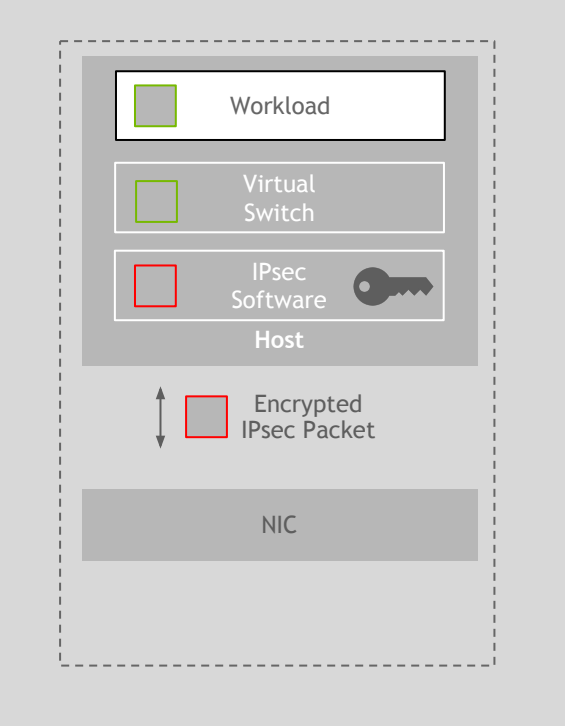
Cipher: AES-GCM 128/256bit keys



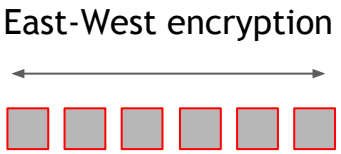


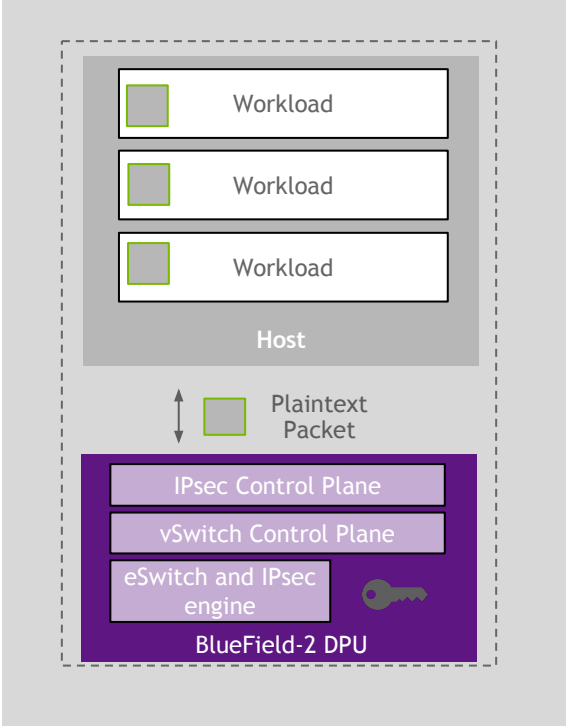
TRANSPARENT IPSEC ENCRYPTION

Encryption/decryption at 100Gb/s bidirectional



Traditional Server
IPsec runs on CPU





DPU Accelerated Server
IPsec and vSwitch on DPU
Inline with other accelerators (tunneling, TLS, etc.)
Cipher: AES-GCM 128/256bit keys
Keys are stored encrypted in hardware
Encrypted RDMA

Plaintext Packet
Encrypted Packet

Control Plane Software on Arm
Acceleration Engine

BLUEFIELD-2 DPU BLOCK DIAGRAM

200 Gbps Ethernet & InfiniBand, NRZ & PAM4 modulation

Powered by ConnectX-6 Dx

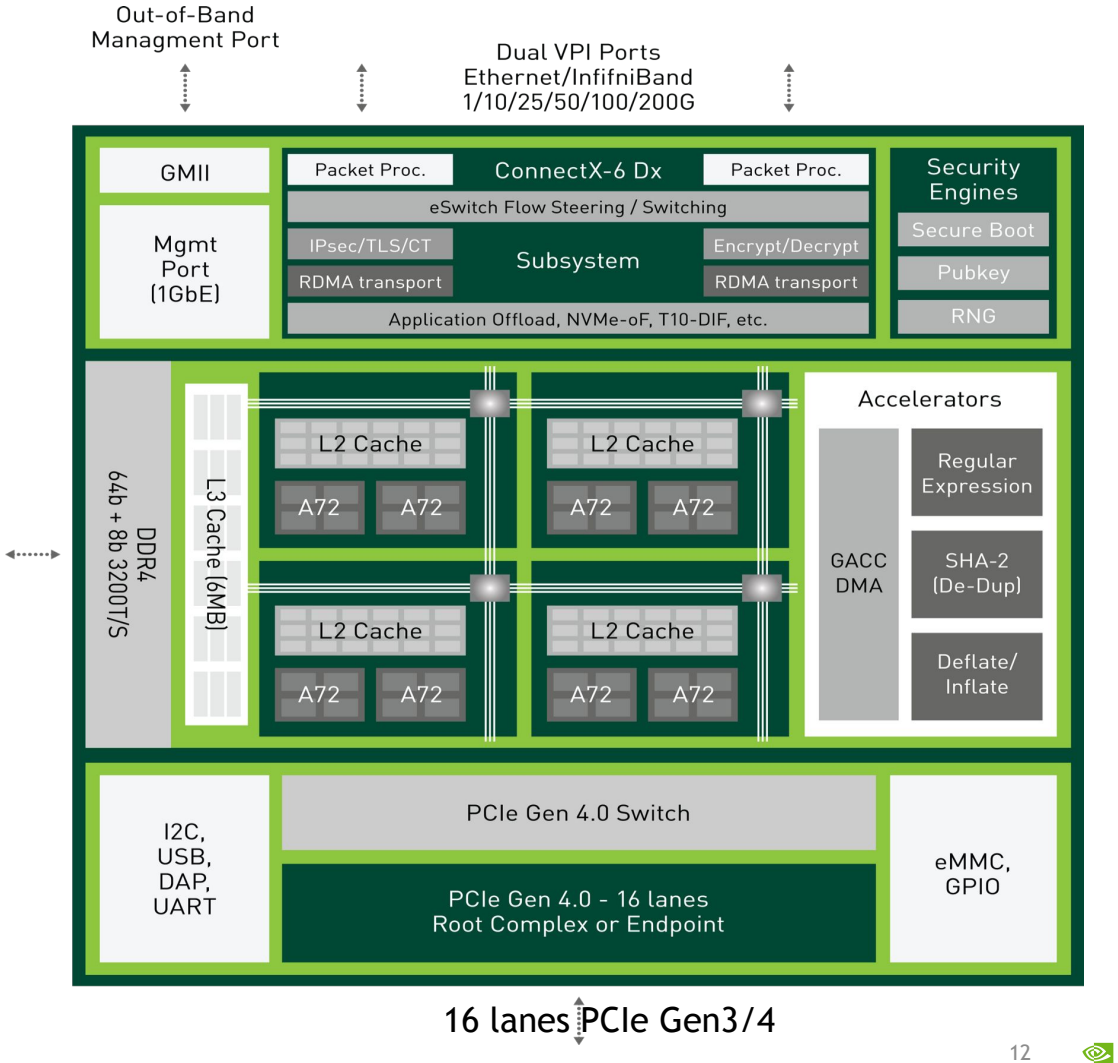
8 ARM A72 CPUs subsystem in a Tile architecture

- 8MB L2 cache, 6MB L3 cache in 4 Tiles
- ARM Frequency up-to 2.5GHz

Fully integrated PCIe switch, 16 bi-furcated Gen4.0

- Root Complex or End Point modes

1GbE Out-of-Band management port



Integrating a DPU into OpenShift

- Motivation
- Starting with RHEL
  a. How it was achieved
  b. Results
- OpenShift Integration
  a. How it was achieved
  b. Results
- Future Work
- Takeaways

Motivation

- Hybrid Cloud Needs
  a. Packets Encryption
  b. Packets Encapsulation
  c. Packets Switching

Secure and Open source RHEL is needed For orchestration, security, stability, and efficiency.

- Separation of dataplane and workload
- De-loading of worker nodes
- 100 Gbps is not feasible without hardware offload

How it was achieved with RHEL

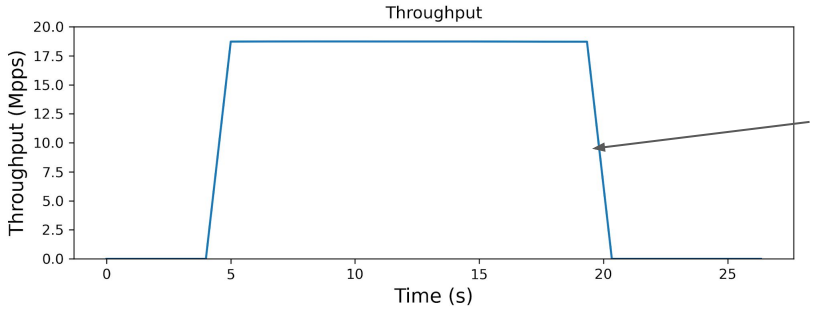
- RHEL8 unmodified on Bluefield-2
- Datapath moved to Bluefield-2
  a. libreswan
  b. Open vSwitch 2.13

RHEL Results

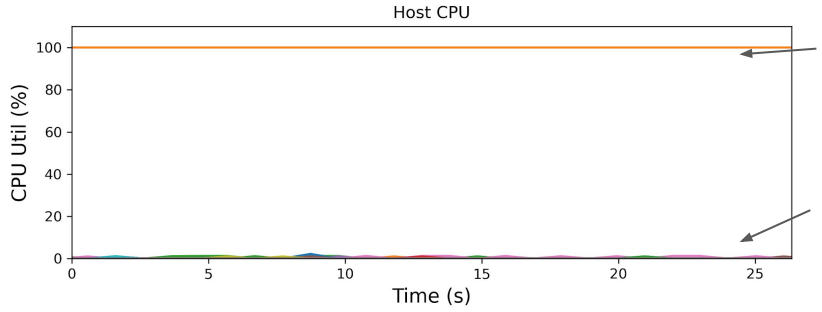
- RHEL Physical-container-physical
  - packets per second (pps)
  - Latency
- RHEL east-west (IPsec)
  - Throughput
  - CPU utilization
- Full Offload Results
  - Encryption, encapsulation, switching

25 Gbps results
and
100 Gbps results

Benchmarks:
Physical-Container-Physical (w/ BlueField-2 OVS offload) 25 Gbps

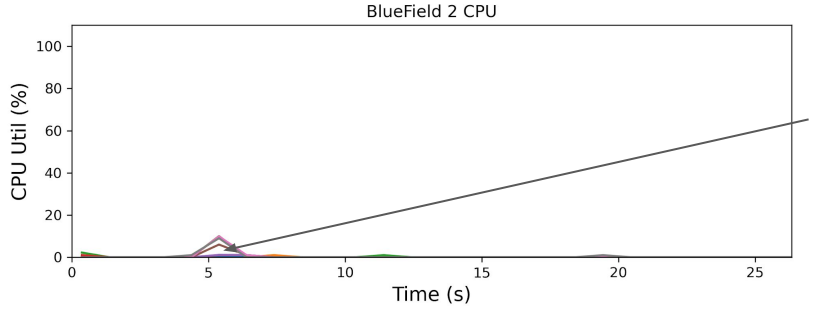


64B packet throughput: 18.7 Mpps
Average latency: 12.9 µs

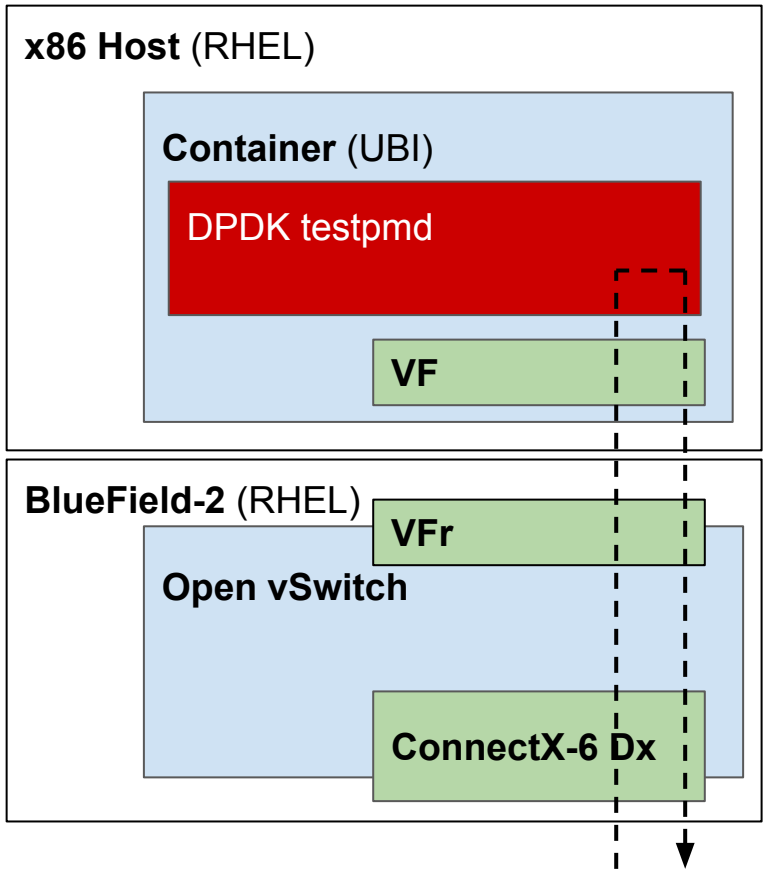


Container (DPDK)

Remaining cores idle



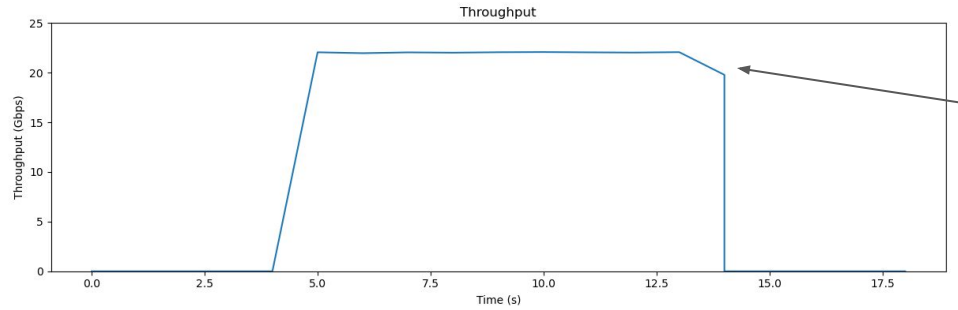
Flow offloaded to ConnectX-6 Dx



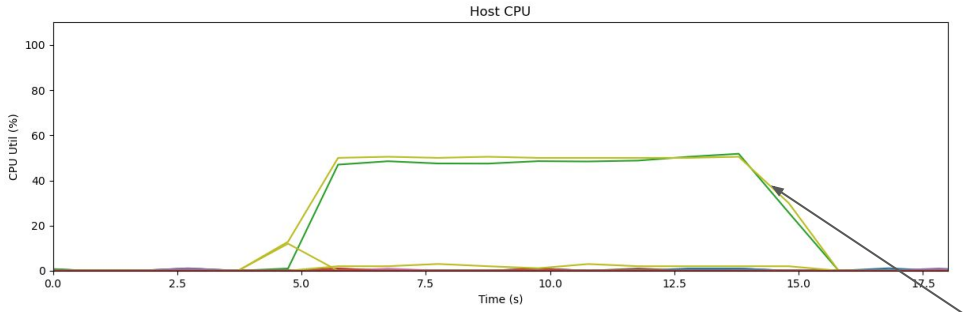
Red Hat

Benchmarks:

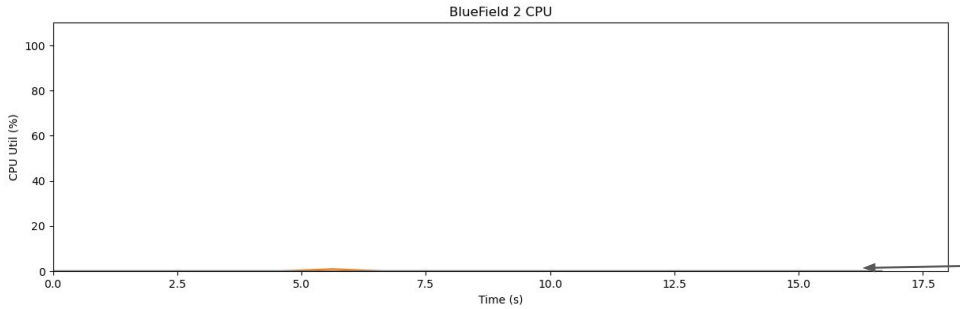
IPsec east-west, (w/ IPsec offloaded) 25 Gbps



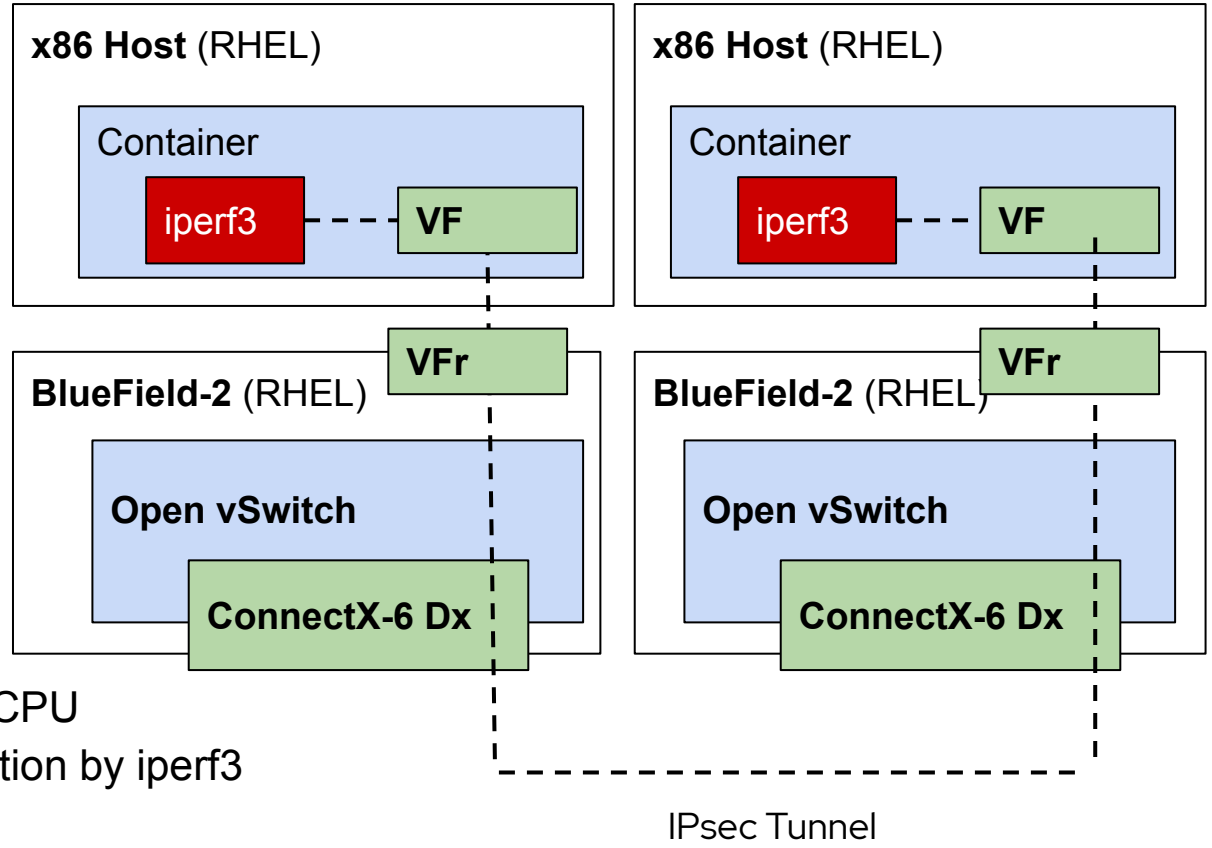
Line rate



Host CPU utilization by iperf3

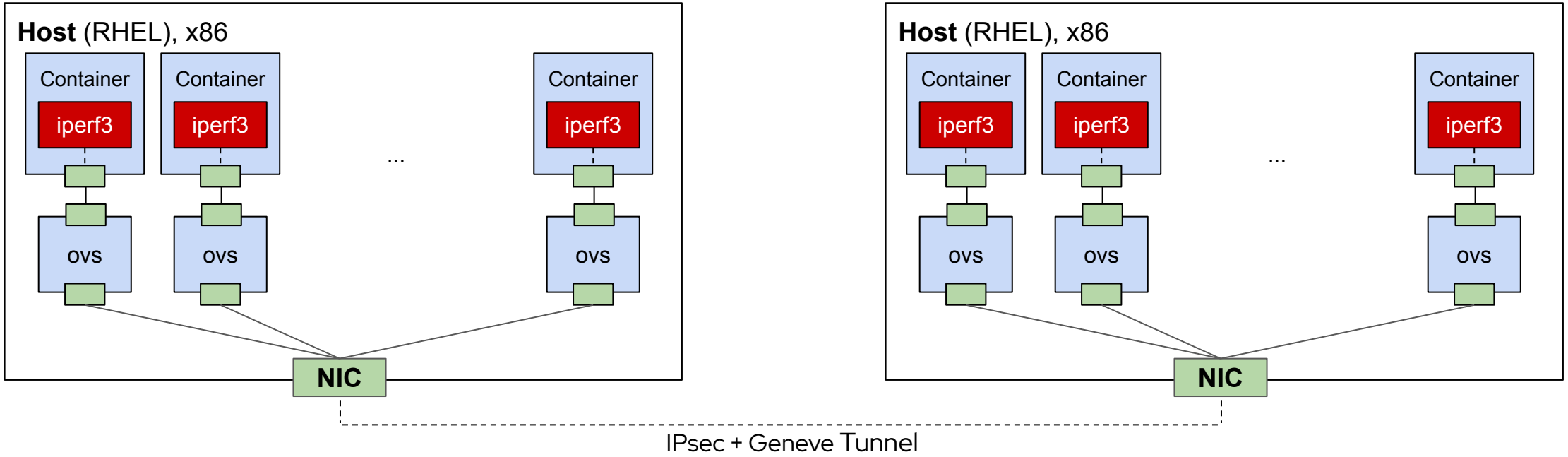


BlueField-2 CPU idle



Red Hat

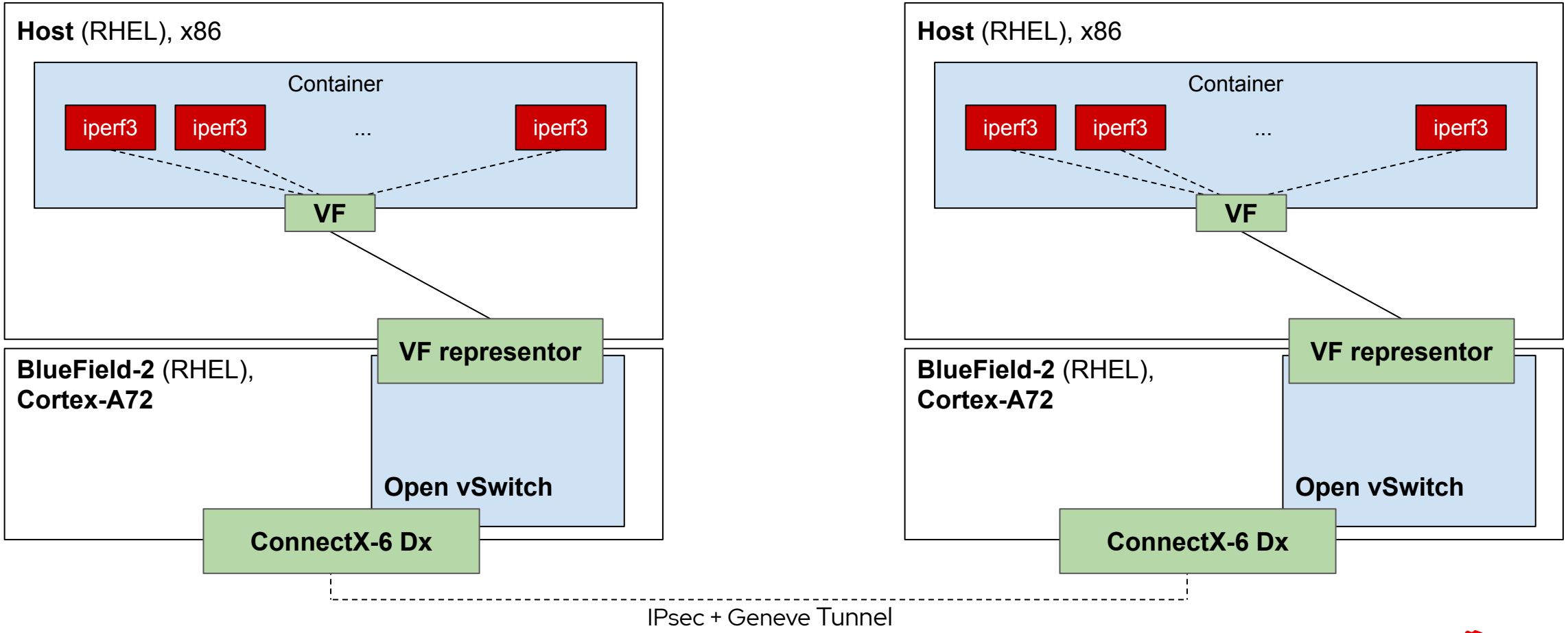
Topology for IPsec + Geneve + OVS benchmark 25 Gbps*: Dataplane on Host (Everything in software)



* 6 IPsec + Geneve tunnels were required to reach line-rate, to use multiple cores on x86

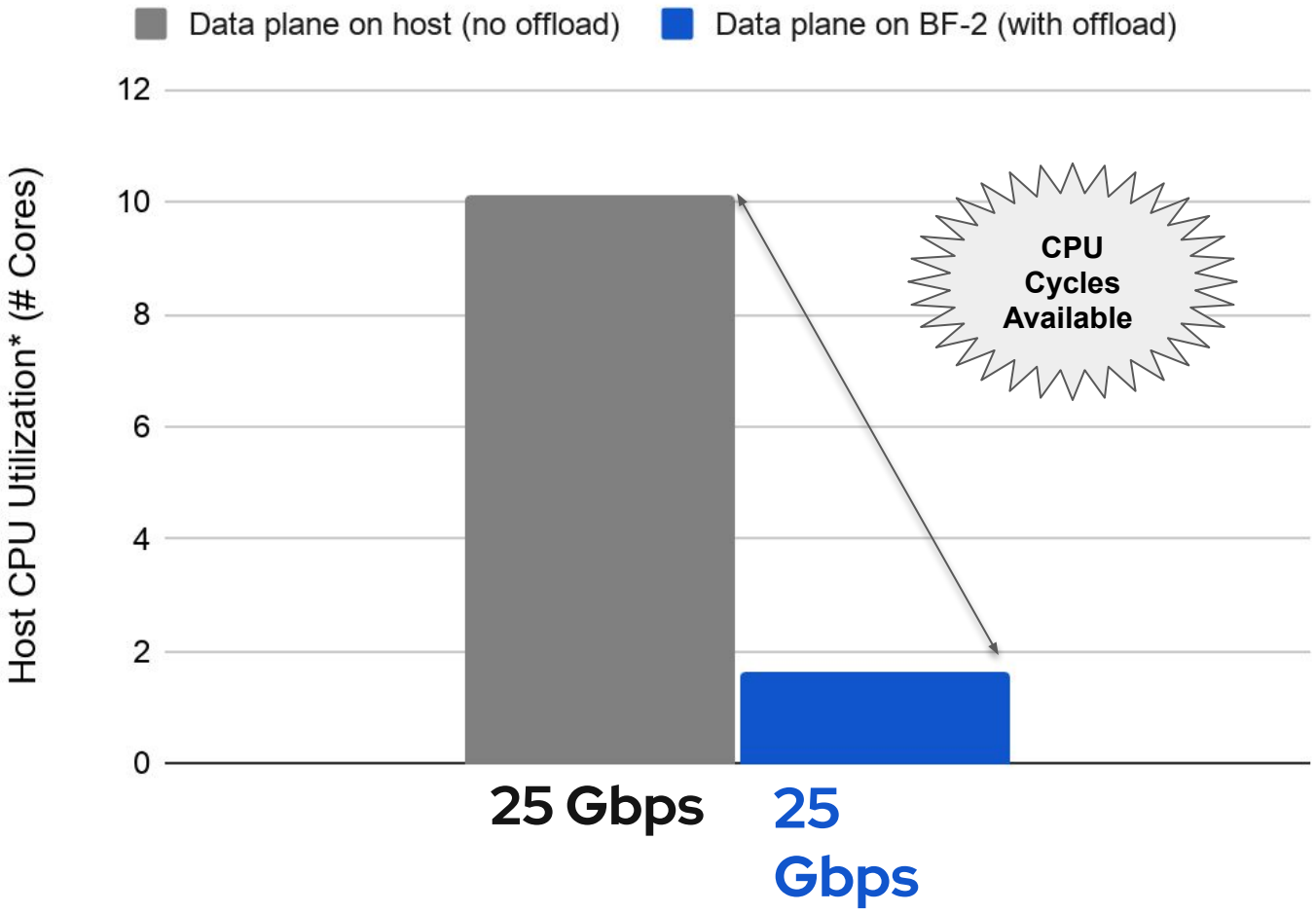
Red Hat

Topology for IPsec + Geneve + OVS benchmark 25 Gbps*: Dataplane on BlueField-2 (IPsec and Geneve offloaded)



* Traffic from all 5 iperf instances through 1 IPsec tunnel

X86 (Motherboard) CPU consumption vs. BlueField-2 HW Offload 25 Gbps

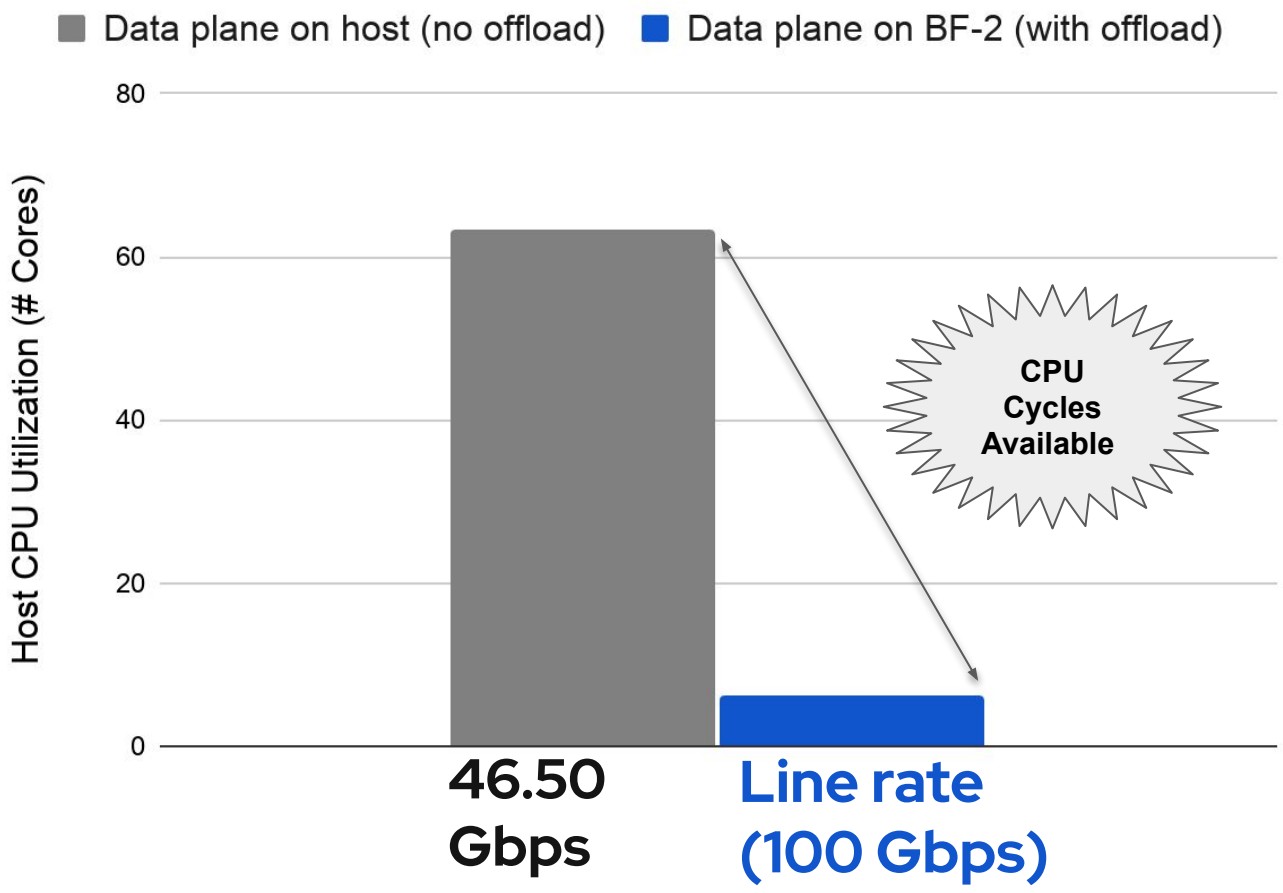


IPsec +
Geneve +
OVS / OVN

Encryption,
Encapsulation,
Switching,
Full HW Offload

* Aggregate CPU utilization (RX node + TX node)

X86 (Motherboard) CPU consumption vs. BlueField-2 HW Offload 100 Gbps

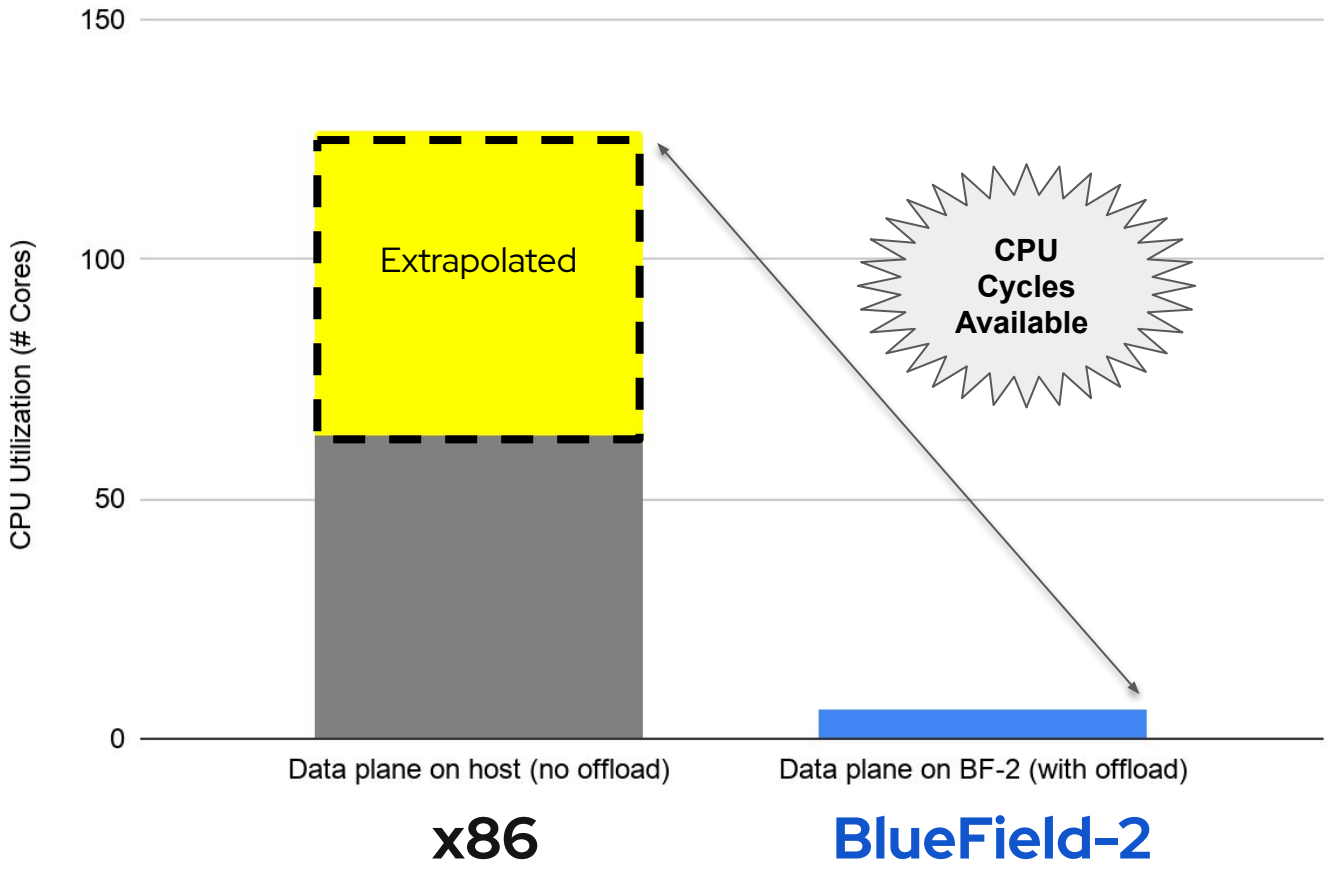


IPsec +
Geneve +
OVS / OVN

Encryption,
Encapsulation,
Switching,
Full HW Offload

* Aggregate CPU utilization (RX node + Tx node)

X86 (Motherboard) CPU consumption vs. BlueField-2 HW Offload 100 Gbps



IPsec +
Geneve +
OVS / OVN

Encryption,
Encapsulation,
Switching,
Full HW Offload

* Aggregate CPU utilization (RX node + Tx node)

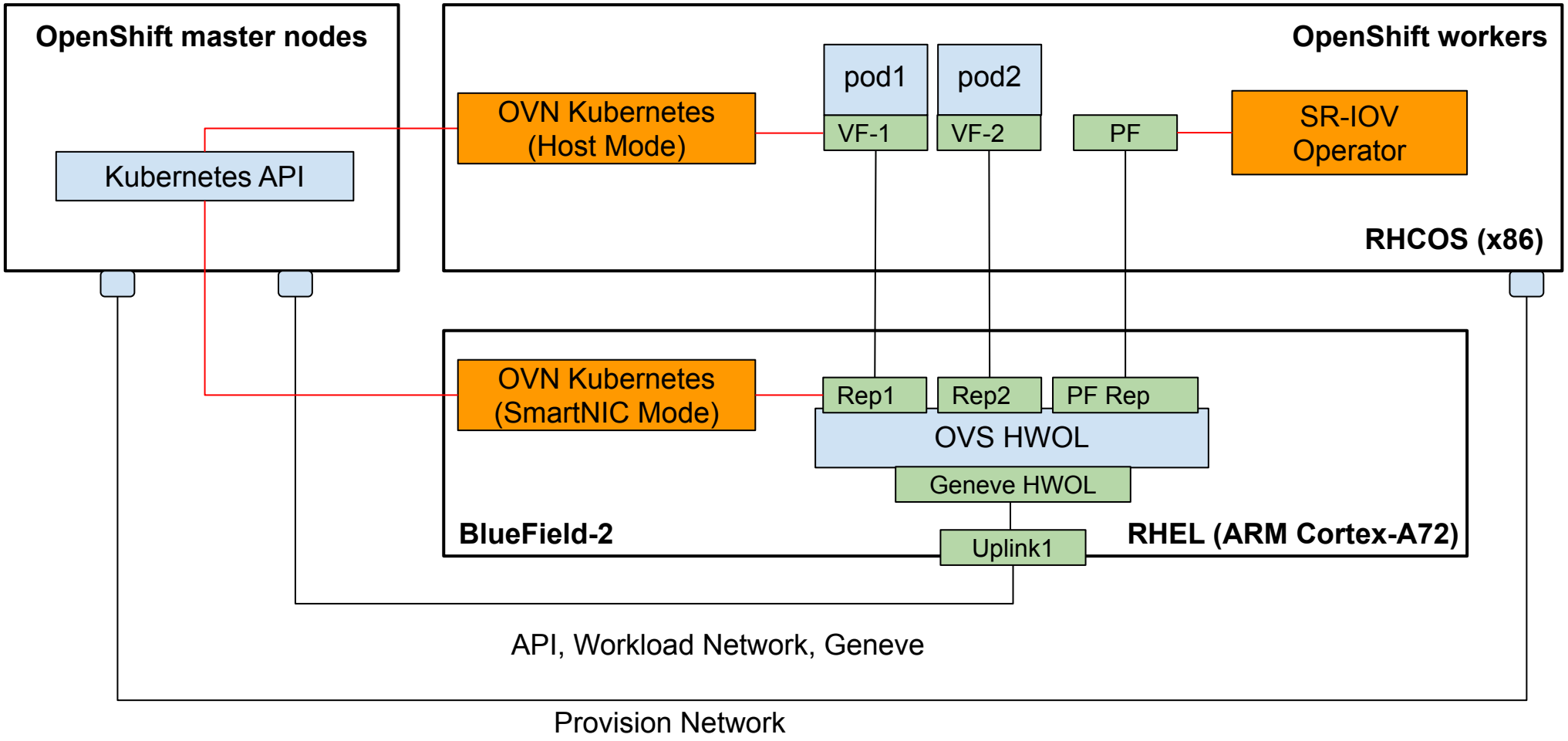
OpenShift Integration

- Kubernetes orchestration between host and BlueField-2
- ovn-kubernetes container running on BlueField-2
- Virtual Function representer on Bluefield-2 sends packets directly to pods
- Hardware offload: IPsec, Geneve, connection tracking

Kubernetes orchestration between host and Bluefield-2

- ovn-k8s-overlay-cni
  - Works with SR-IOV operator
  - Places Virtual Function (VF) in worker pod
- ovnkube-node runs on x86 worker node and Bluefield-2
  - Communication via kubernetes APIs
  - Attaches Virtual Function representer (VFr) to OVS on Bluefield-2

Integrating a DPU into OpenShift



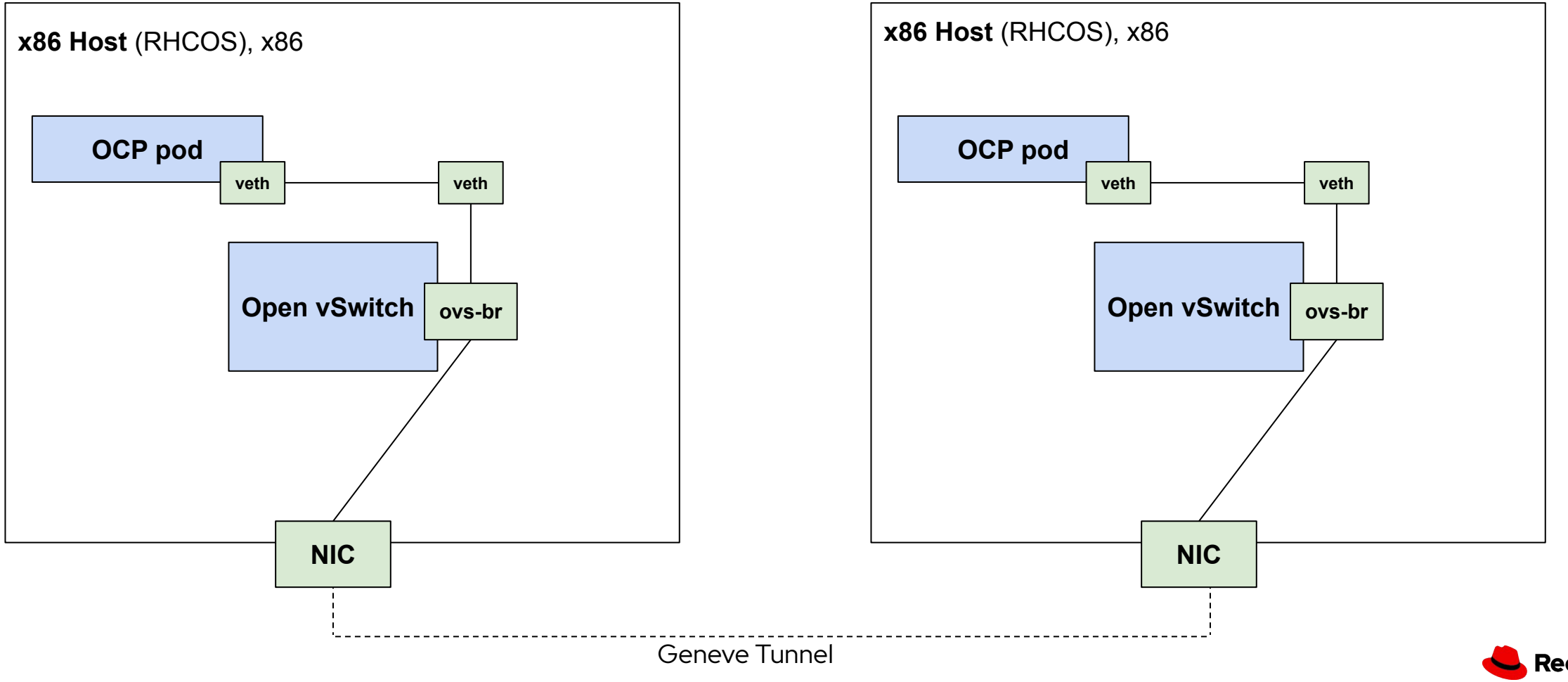
Red Hat

OpenShift Results

- OpenShift east-west host CPU utilization
  - ovn-kubernetes on x86 host
  - ovn-kubernetes on BlueField-2

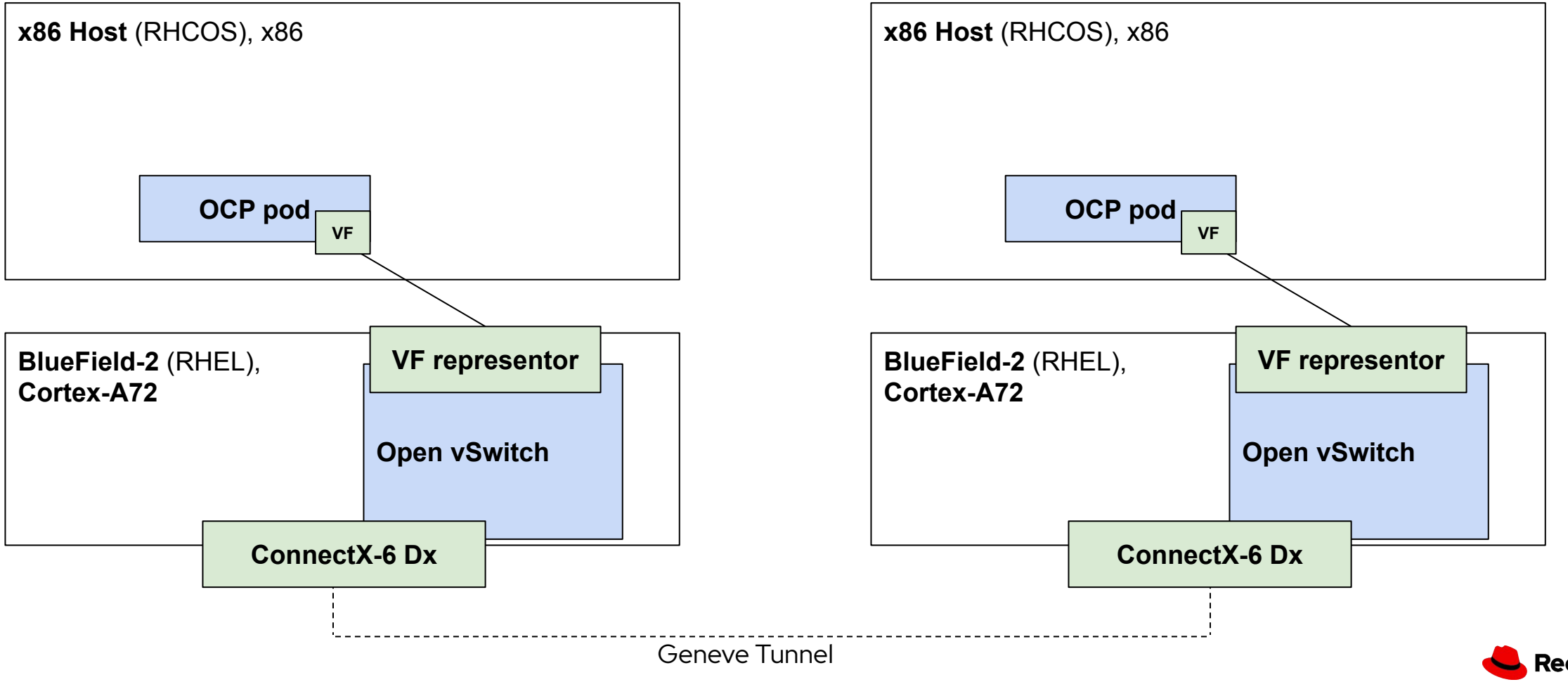
OpenShift east-west:

ovn-kubernetes on x86 host

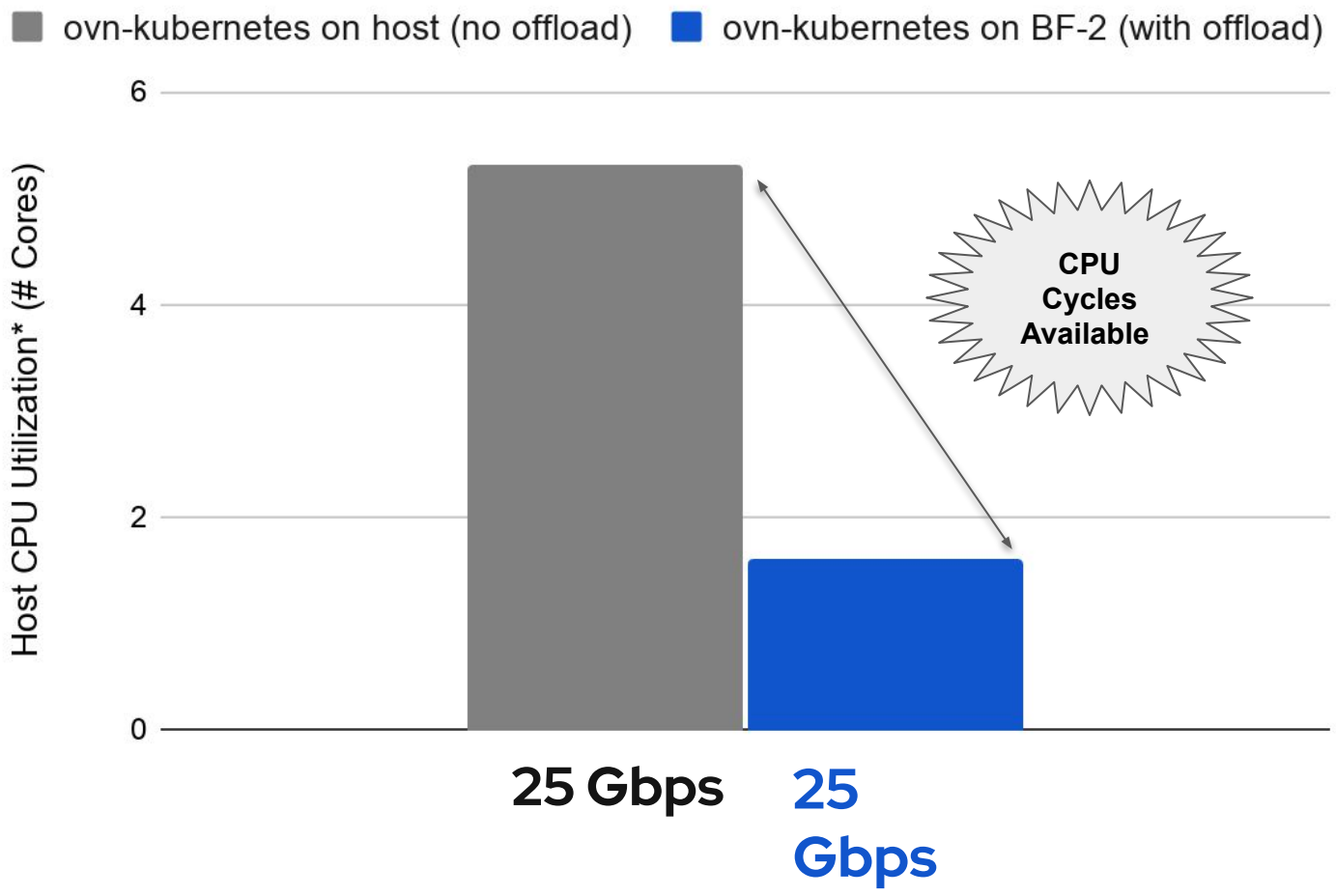


OpenShift east-west:

ovn-kubernetes on BlueField-2



X86 (Motherboard) CPU consumption vs. BlueField-2 HW Offload 25 Gbps



Geneve +
OVS / OVN

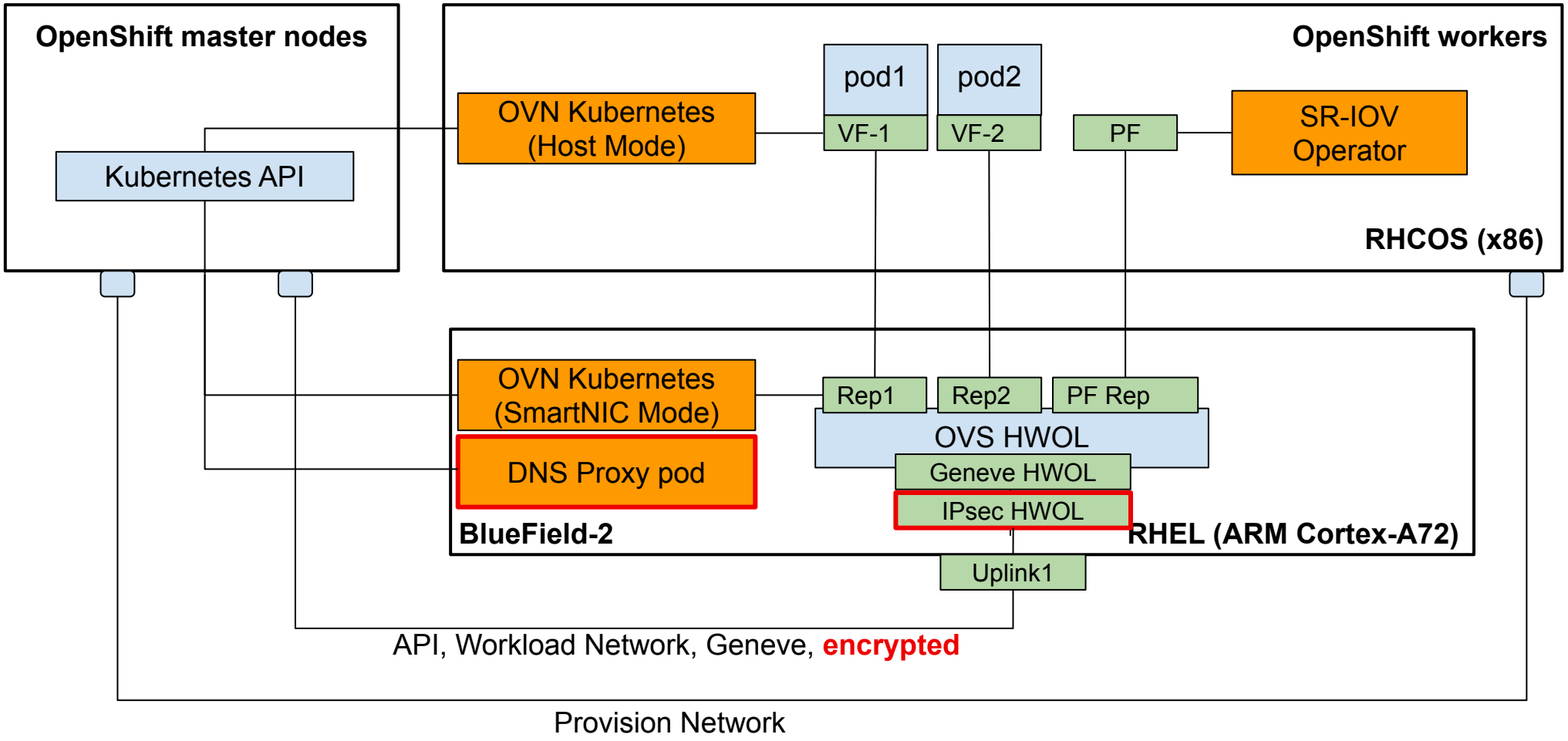
Encapsulation,
Switching,
Full HW Offload

* Aggregate CPU utilization (RX node + TX node)

Future Work

- BlueField-2 as a first class member of OpenShift cluster
- OpenShift IPsec (encrypted east-west)
- Kubernetes shared gateway mode
- OpenShift DNS proxy pods to BlueField-2

Future Work: OpenShift Integration



RED denotes future work items

Red Hat

Takeaways

- BlueField-2 is by itself is a paradigm shift
  a. Unmatched performance and reliability
- Red Hat Openshift, Openstack and RHEL provide
  a. Orchestration
  b. Stability
  c. Performance
  d. Reliability
  e. Security
  f. Support
- Red Hat + Nvidia will change the hybrid cloud and data centers forever

Thank you

Red Hat is the world's leading provider of enterprise open source software solutions. Award-winning support, training, and consulting services make Red Hat a trusted adviser to the Fortune 500.

linkedin.com/company/red-hat

youtube.com/user/RedHatVideos

facebook.com/redhatinc

twitter.com/RedHat

Red Hat

BACKUP SLIDES

OVS/OVN flow offload

```
[root@bluefield-soc ~]# ovs-appctl dpctl/dump-flows type=offloaded
tunnel(tun_id=0x0,src=10.1.0.2,dst=10.1.0.1,tp_dst=6081,flags(+key)),recirc
_id(0),in_port(3),eth(src=8a:ab:a4:11:de:0d,dst=0e:01:fa:c0:32:22),eth_type
(0x0800),ipv4(frag=no), packets:34639, bytes:1802336, used:0.000s,
actions:2
--
```

Geneve tunnel match

Output to VF representor

OVS/OVN flow offload, Linux traffic control hardware offload

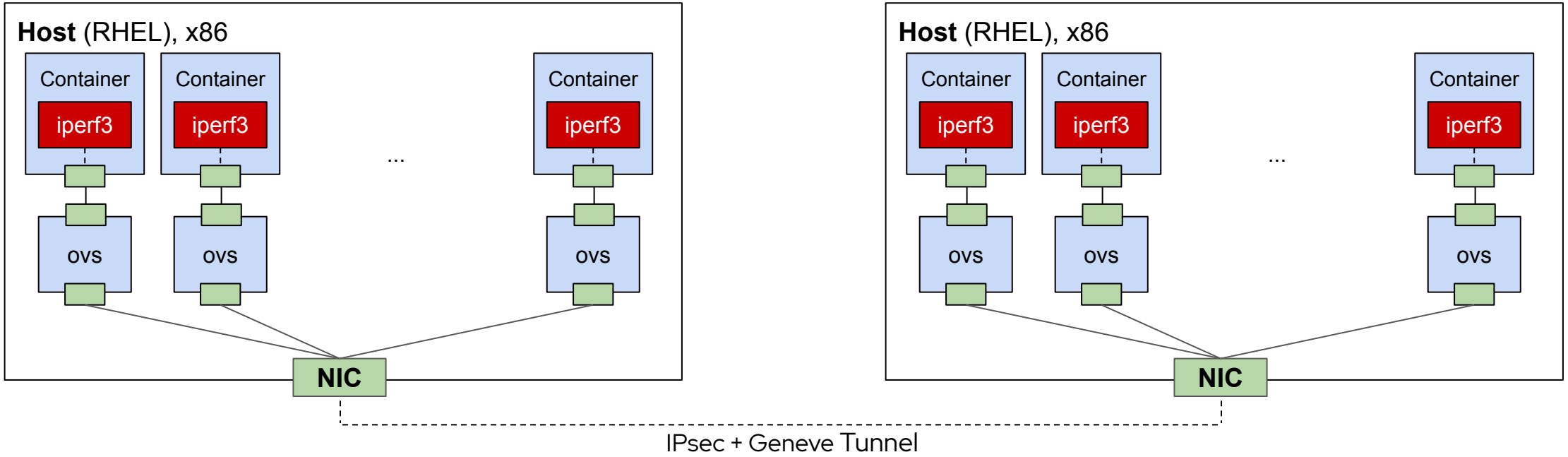
```
# tc filter show dev genev_sys_6081 ingress
[..]
  enc_dst_ip 10.0.0.3
  enc_src_ip 10.0.0.1
  enc_dst_port 6081
        action order 2: mirred (Egress Redirect to device eth0) stolen
--
pod-to-pod (east-west) traffic
```

Terminalizer

```
[fdangelo@ibm-p8-kvm-05-fsp ~]$
```

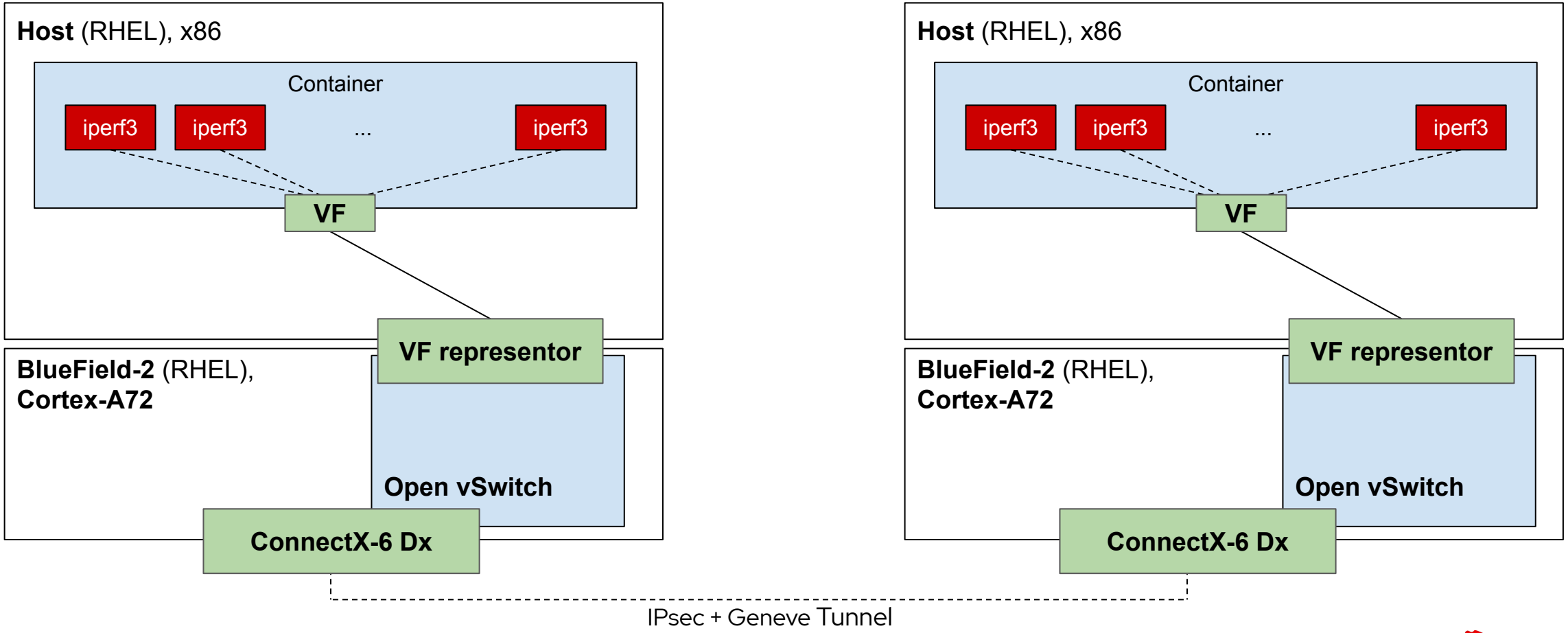

Red Hat

Topology for IPsec + Geneve + OVS benchmark 100 Gbps*: Dataplane on host (Everything in software)



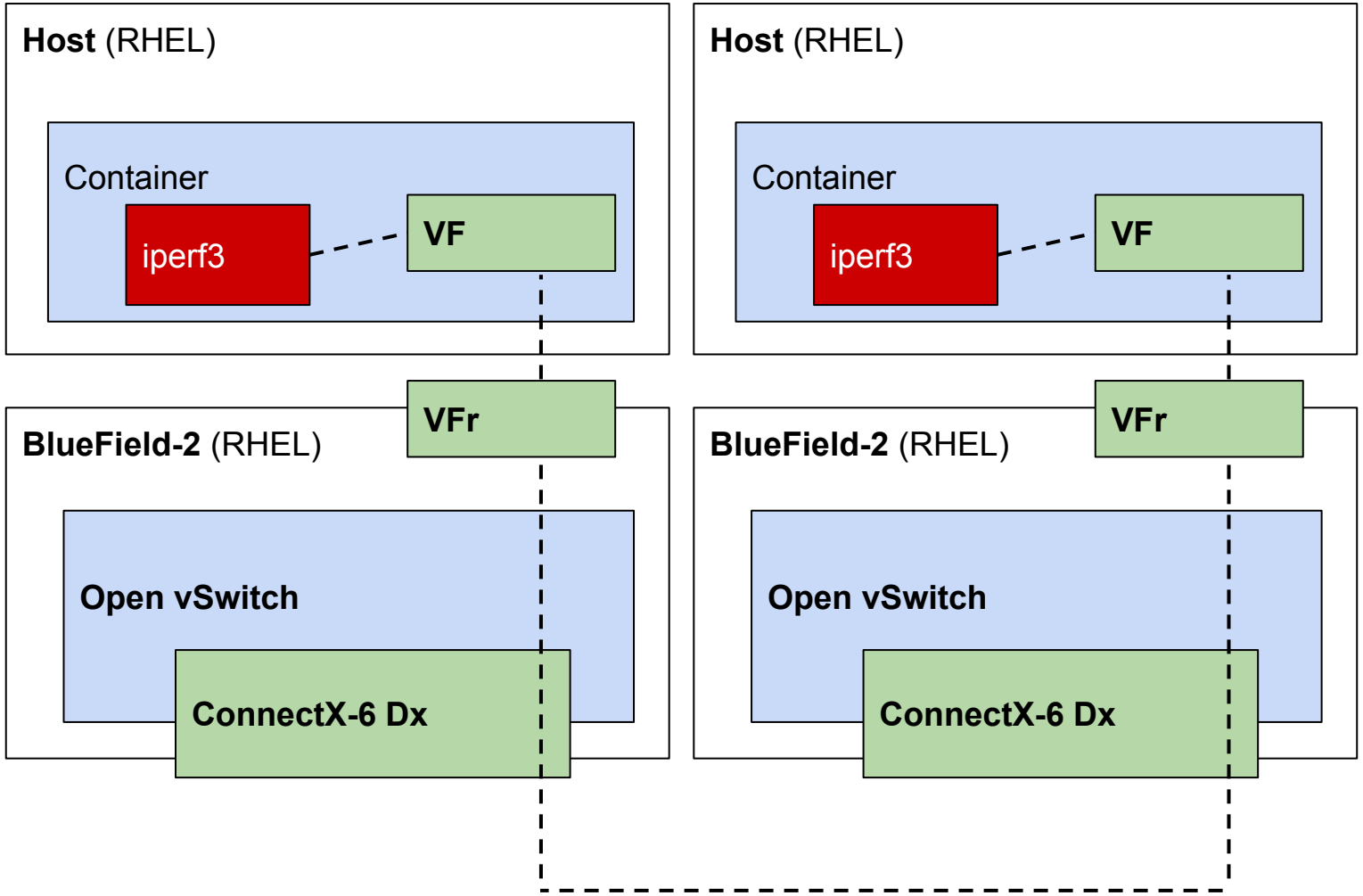
* 100 IPsec + Geneve tunnels were required to reach line-rate, to use multiple cores on x86

Topology for IPsec + Geneve + OVS benchmark 100 Gbps*: Dataplane on BlueField-2 (IPsec and Geneve offloaded)



* Traffic from all 100 iperf instances through 1 IPsec tunnel

RHEL Benchmarking Setup

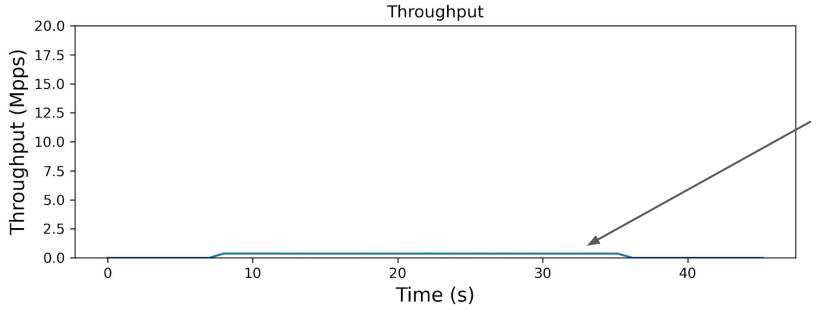


Red Hat

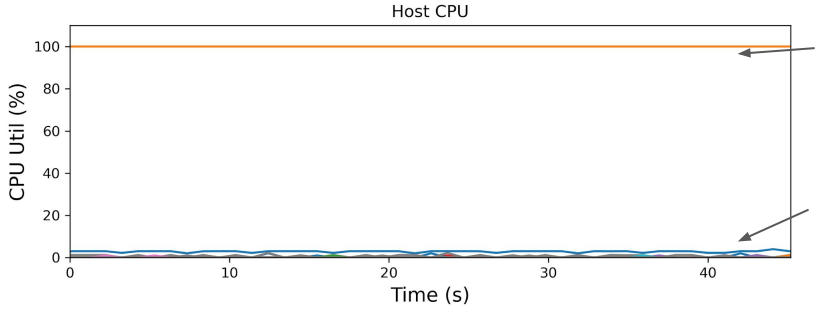
RHEL IPsec CPU Utilization Results

- x86 datapath
  - Line rate, 25 Gbps and 100 Gbps
  - CPU utilization
- Bluefield-2 datapath
  - Line rate, 25 Gbps and 100 Gbps
  - CPU utilization

Benchmarks: Physical-Container-Physical (w/o ConnectX-6 Dx offload)

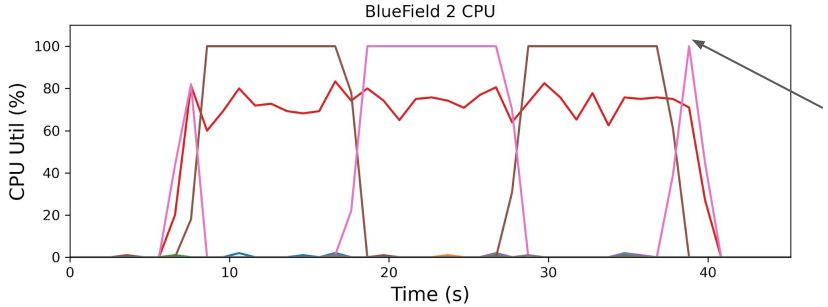


64B packet throughput: 350 Kpps
Average latency: 18.2 µs

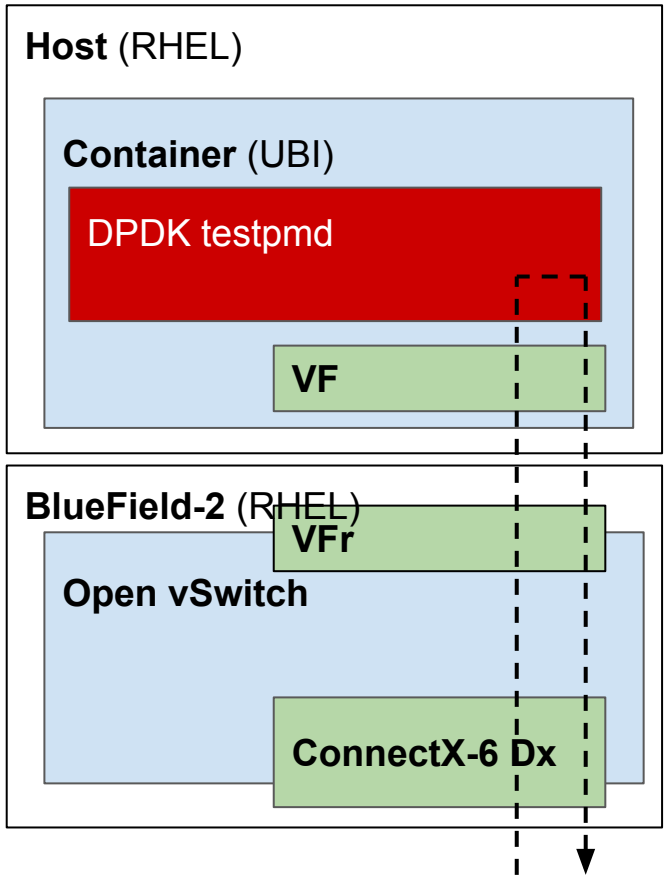


Container (DPDK)

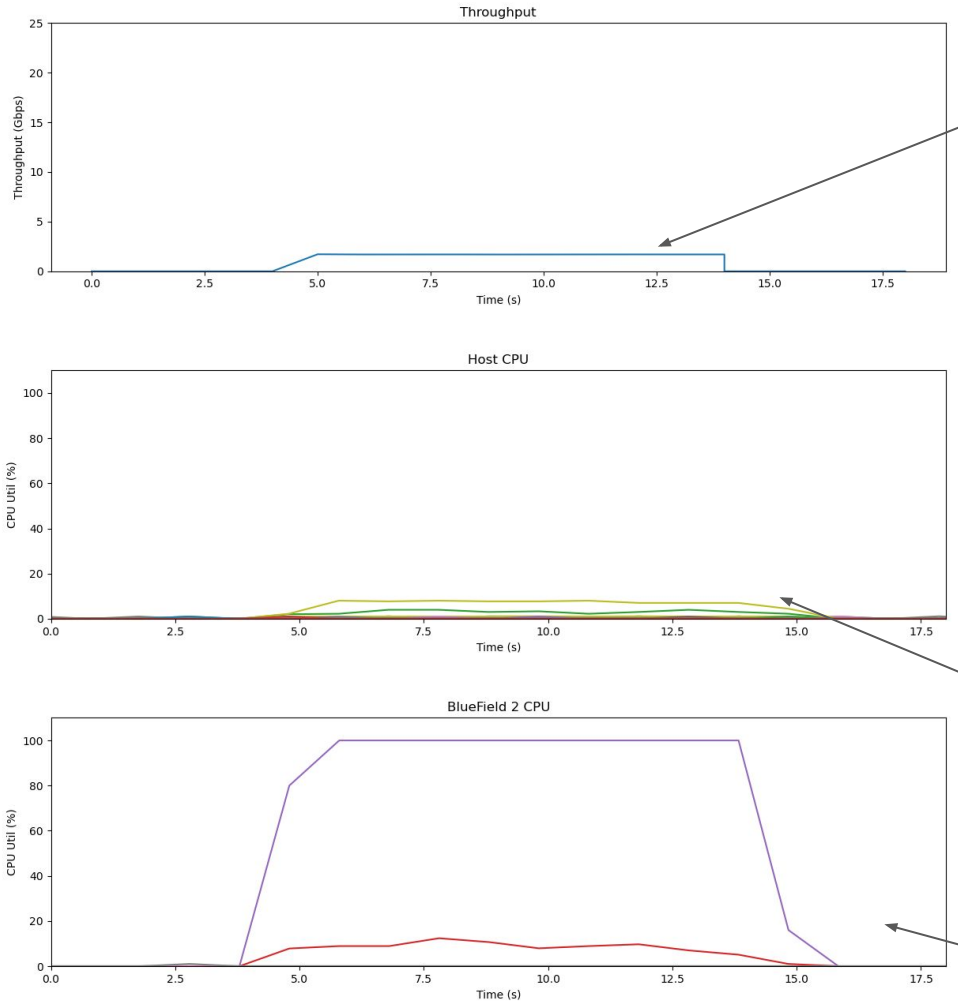
Remaining cores idle



OvS software Datapath



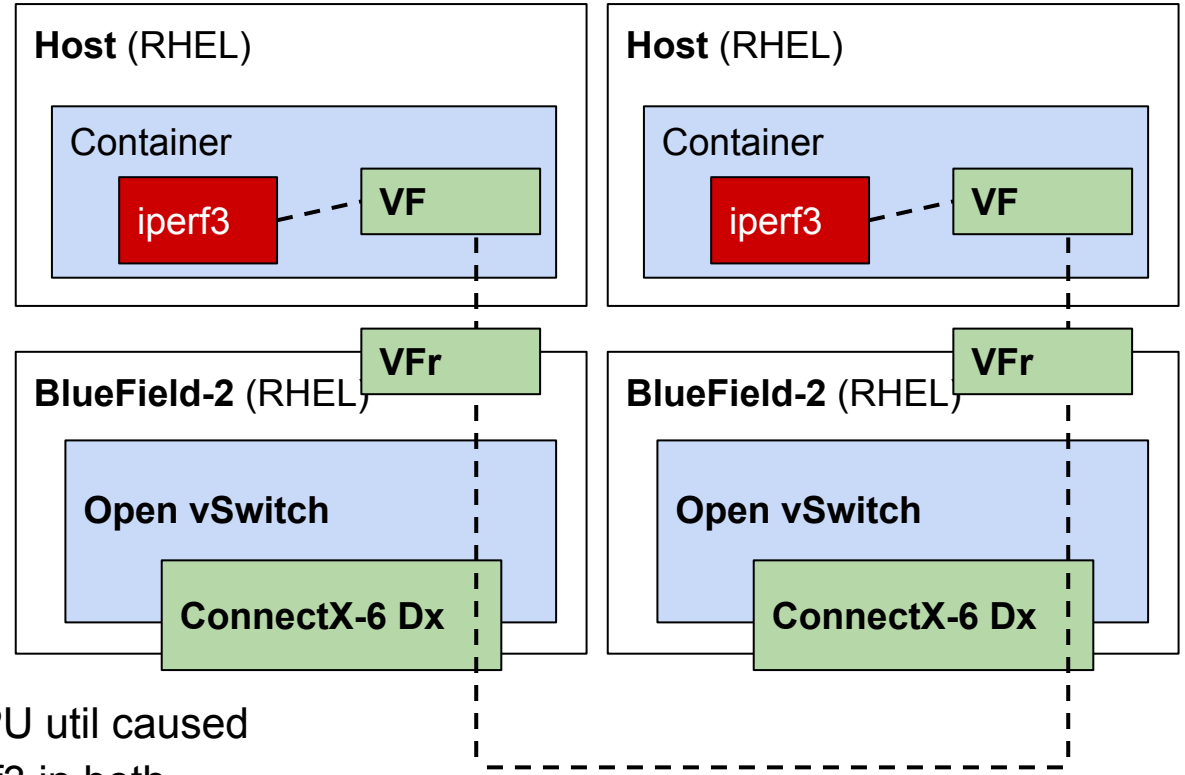
Benchmarks: IPsec east-west, (w/o IPsec offloaded)



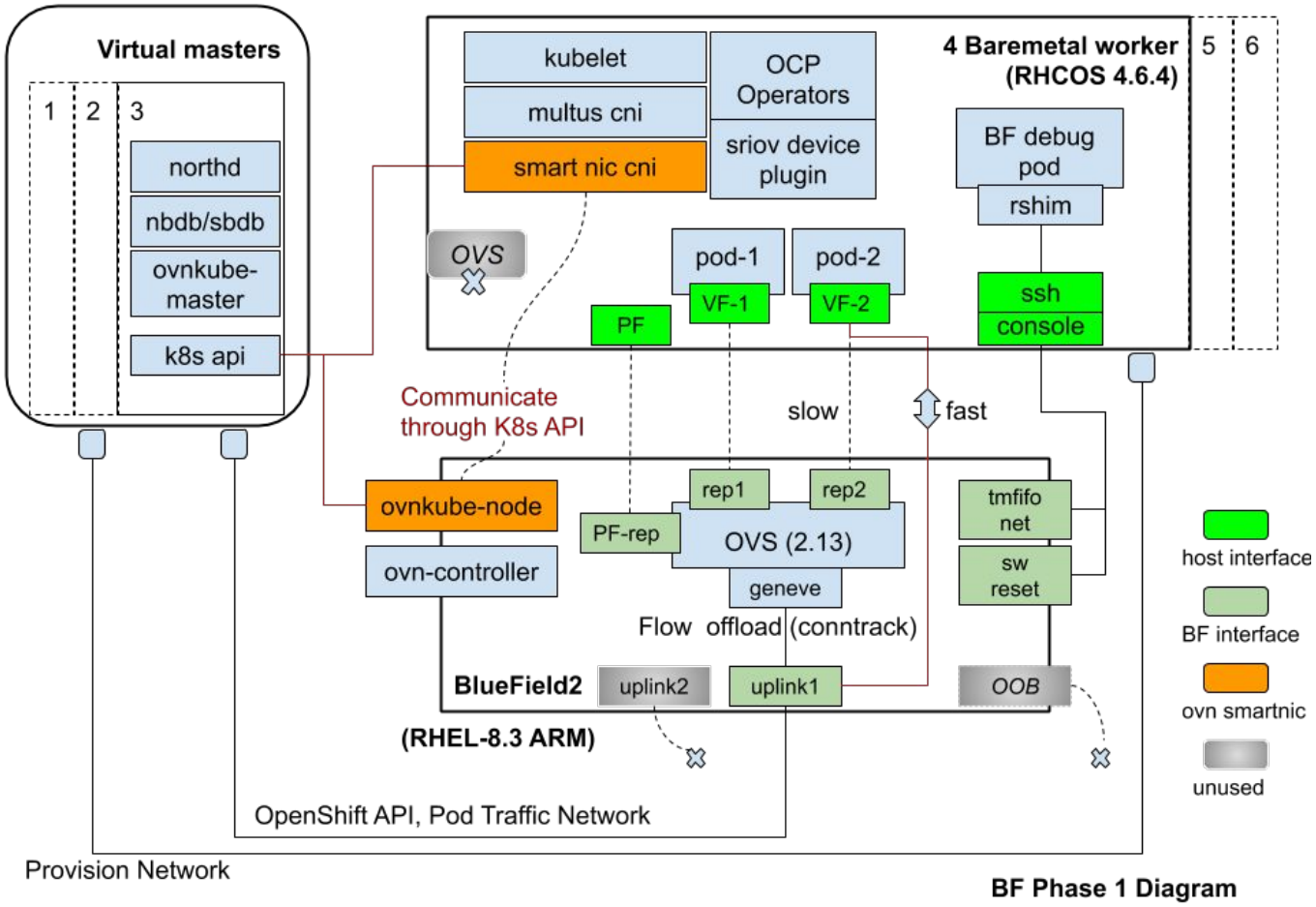
1.6 Gbps

x86 CPU util caused by iperf3 in both userspace and kernel

BlueField-2 CPU bottleneck

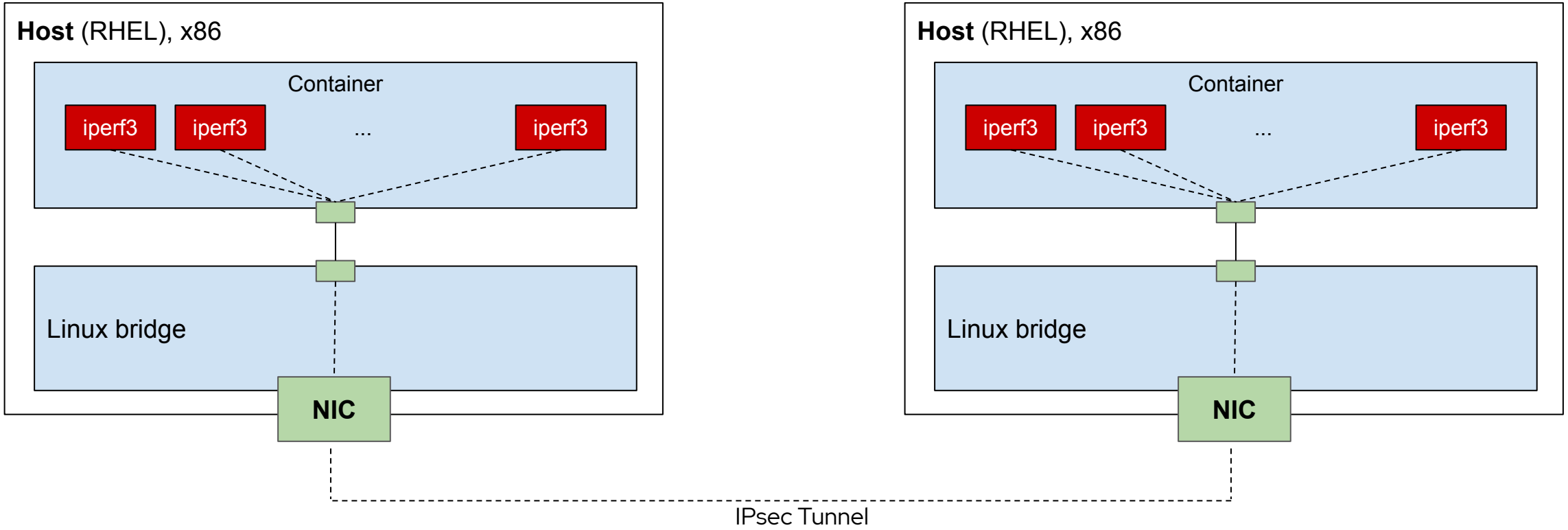


Integrating a SmartNIC into OpenShift



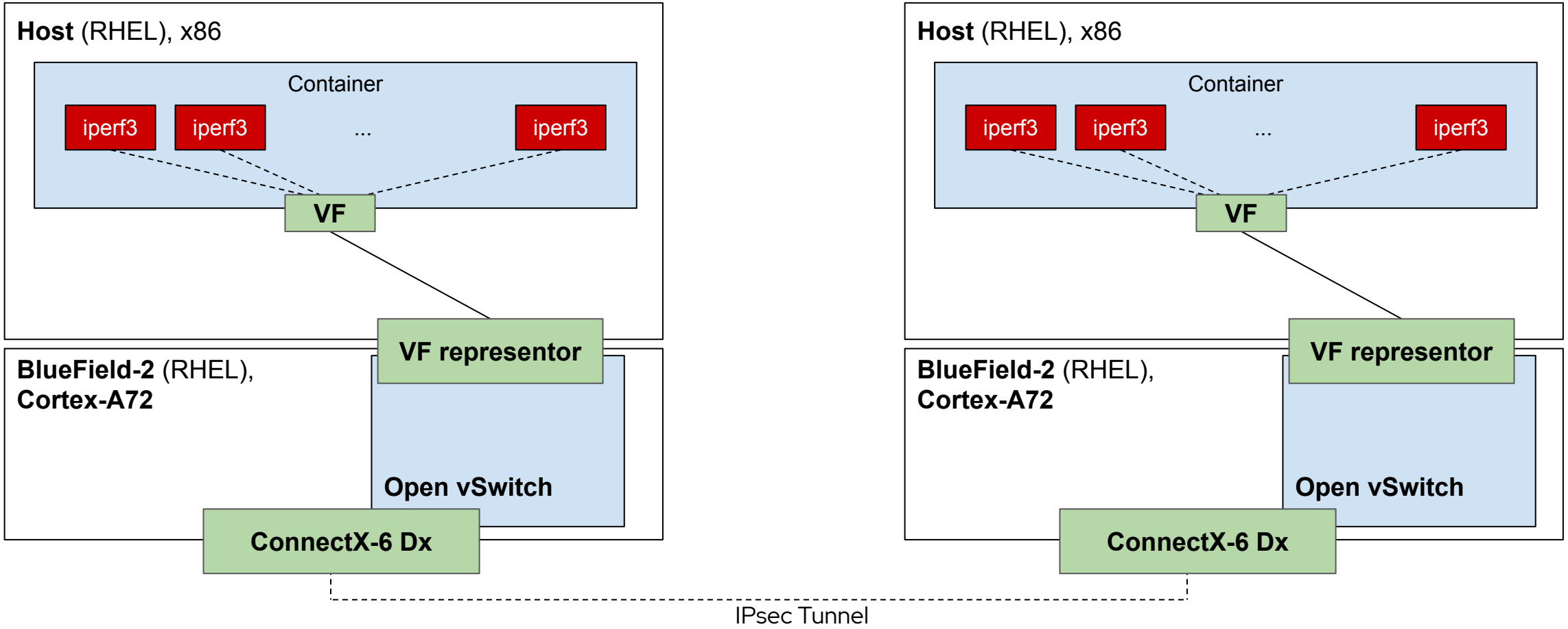
BF Phase 1 Diagram

Topology for IPsec benchmark 25 Gbps*: Dataplane on host (Everything in software)



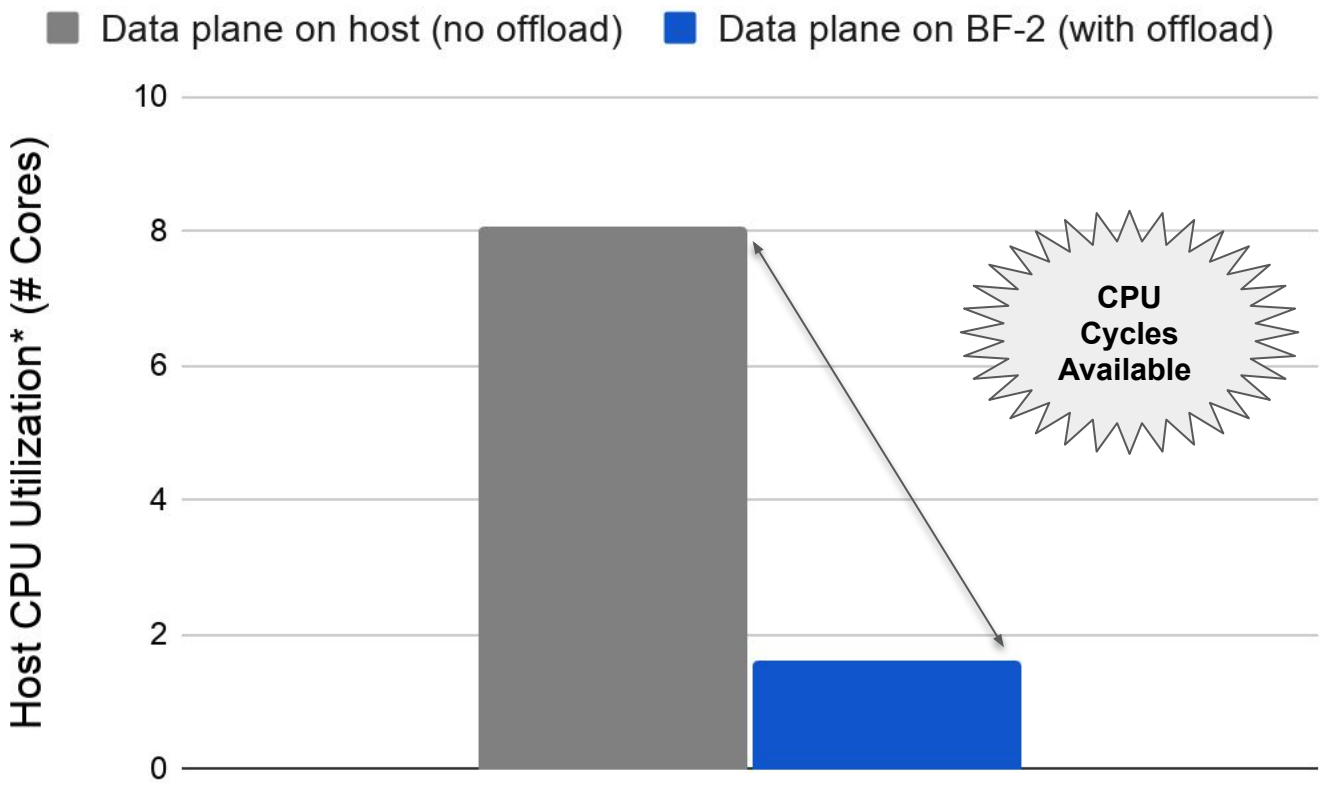
* 5 IPsec tunnels were required to reach line-rate, to use multiple cores on x86

Topology for IPsec benchmark 25 Gbps*: Dataplane on BlueField-2 (IPsec offloaded)



* Traffic from all 5 iperf instsances passed through 1 IPsec tunnel

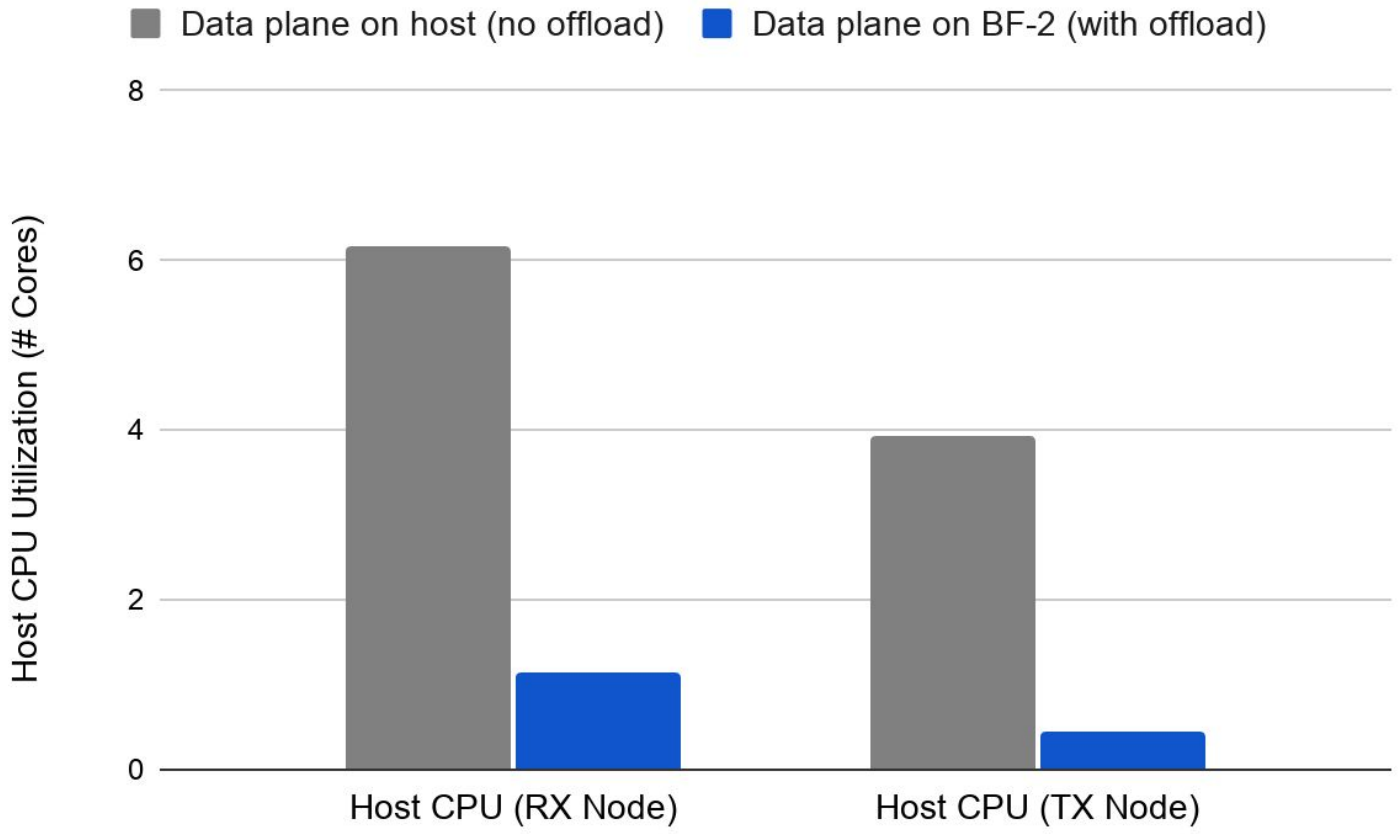
X86 (Motherboard) CPU consumption vs. BlueField-2 HW Offload 25Gbps



IPsec

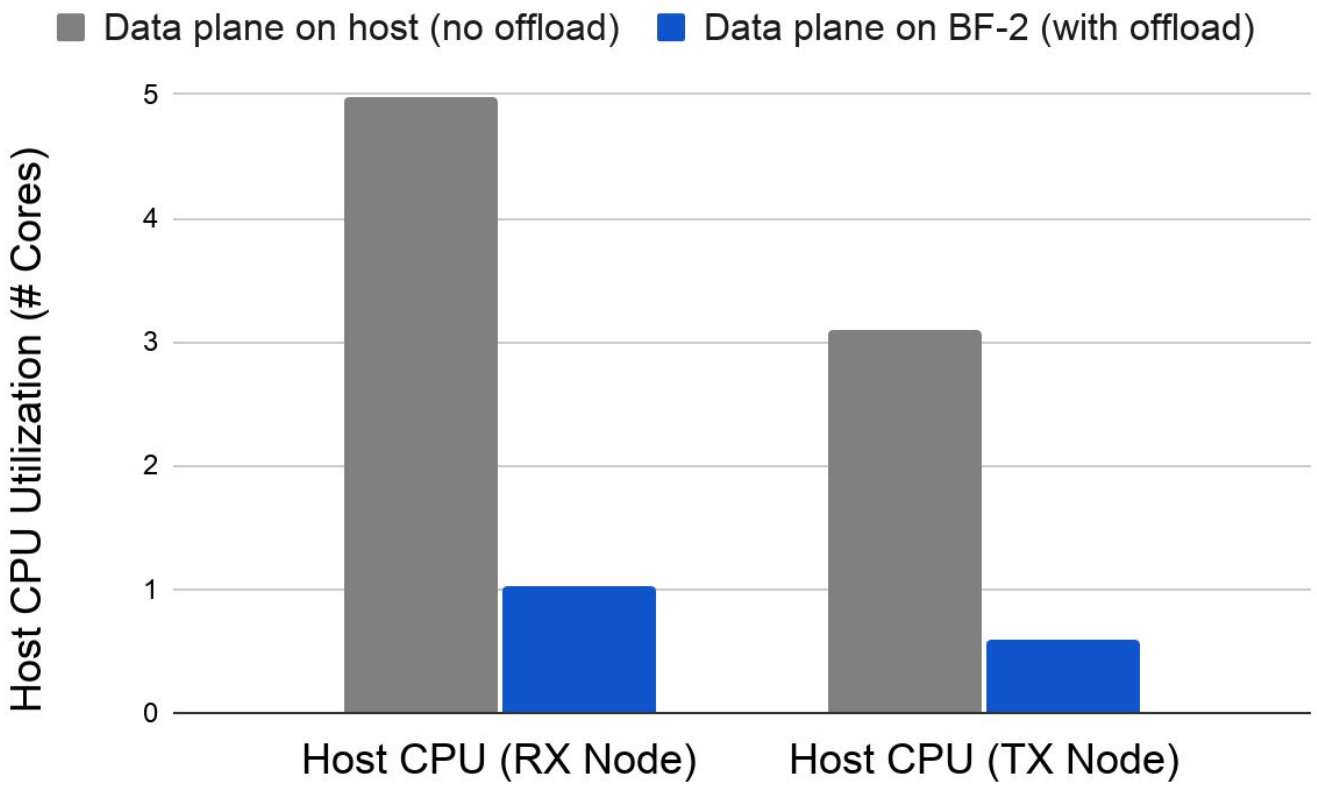
Encryption,

* Aggregate CPU utilization (RX node + TX node)



Benchmarks: IPsec + Geneve + OVS east-west, side-by-side CPU util,
25 Gbps w/ dataplane on x86

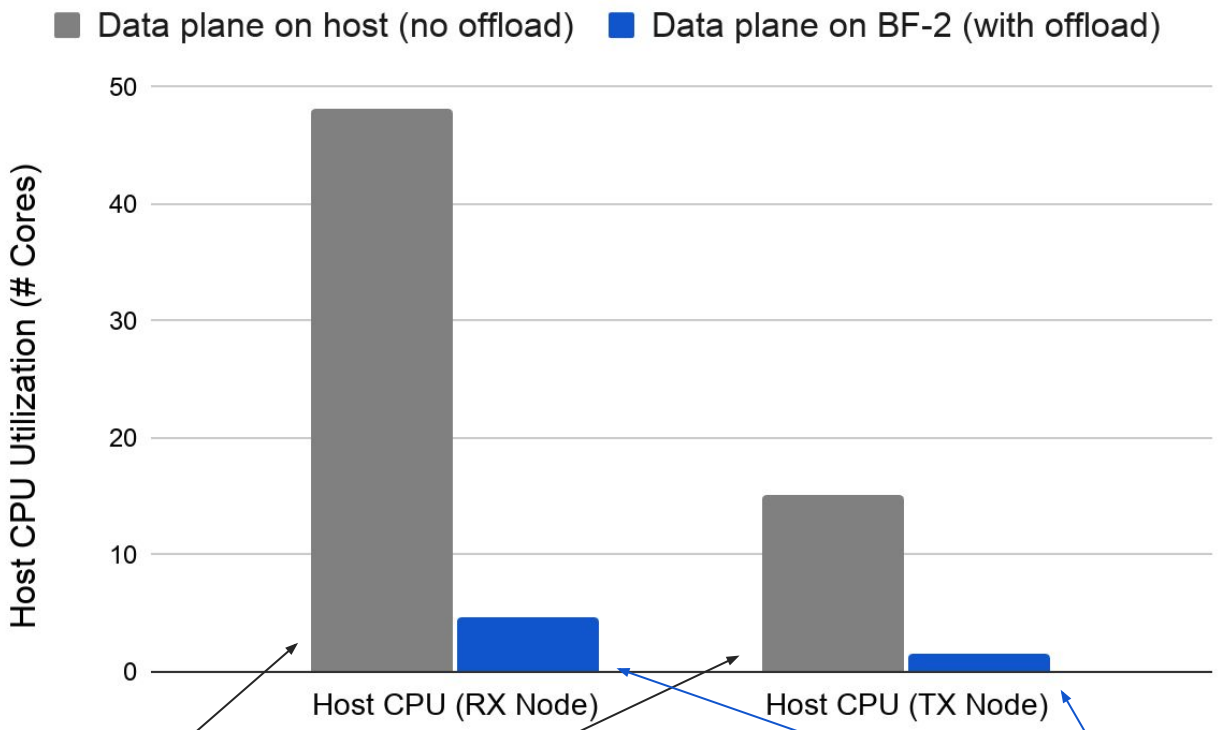
vs 25 Gbps w/ dataplane on BlueField-2.



Benchmarks: IPsec east-west, side-by-side CPU utilization, 25 Gbps w/ dataplane on x86

25 Gbps w/ dataplane on BlueField-2.

X86 (Motherboard) CPU consumption vs. BlueField-2 HW Offload 100G



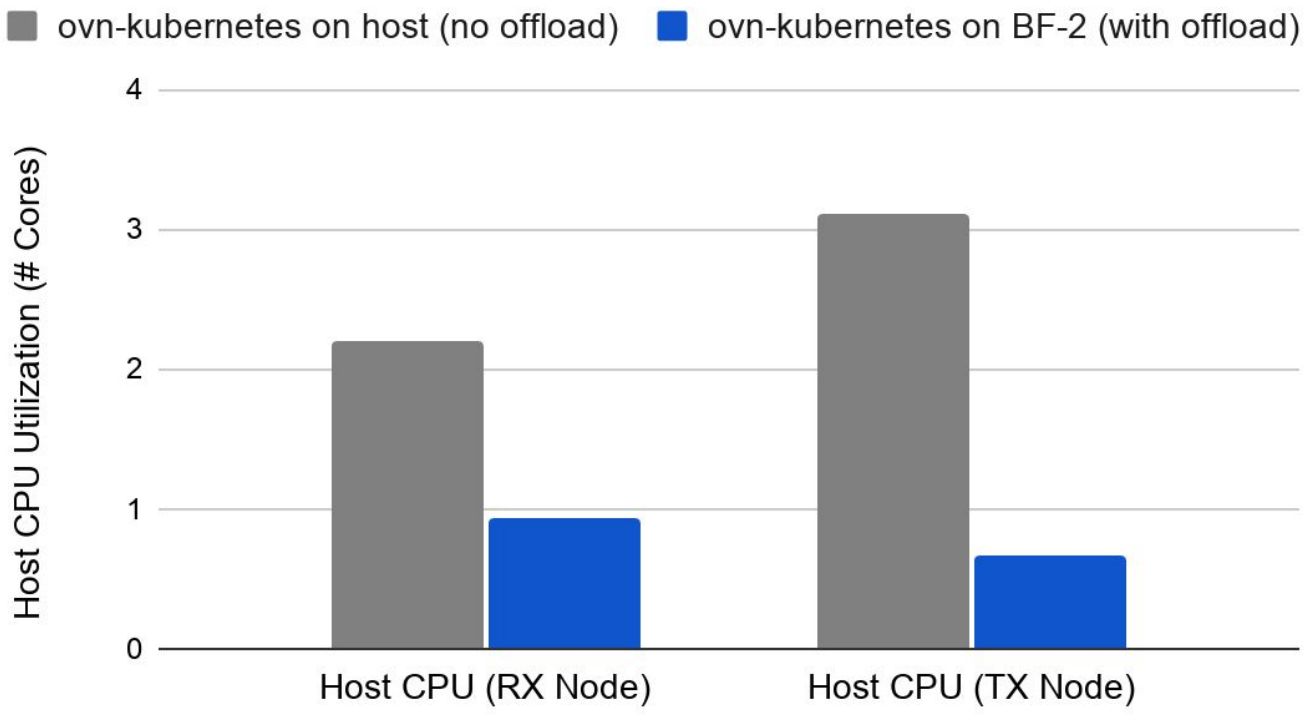
46.50 Gbps

Line rate (100 Gbps)

IPsec +
Geneve +
OVS / OVN

Encryption,
Encapsulation,
Switching,
Full HW Offload

Red Hat



OpenShift: Geneve east-west, side-by-side CPU util, 25 Gbps w/ ovn-kubernetes on host

vs 25 Gbps w/ ovn-kubernetes on BlueField-2.