



Open vSwitch

Mark Kavanagh / Tarek Radi

Intel Corporation

Optimizing TCP Workloads in an OvS-based NFV  
Deployment

# Legal Disclaimer

## **General Disclaimer:**

© Copyright 2016 Intel Corporation. All rights reserved. Intel, the Intel logo, Intel Inside, the Intel Inside logo, Intel. Experience What's Inside are trademarks of Intel. Corporation in the U.S. and/or other countries. \*Other names and brands may be claimed as the property of others.

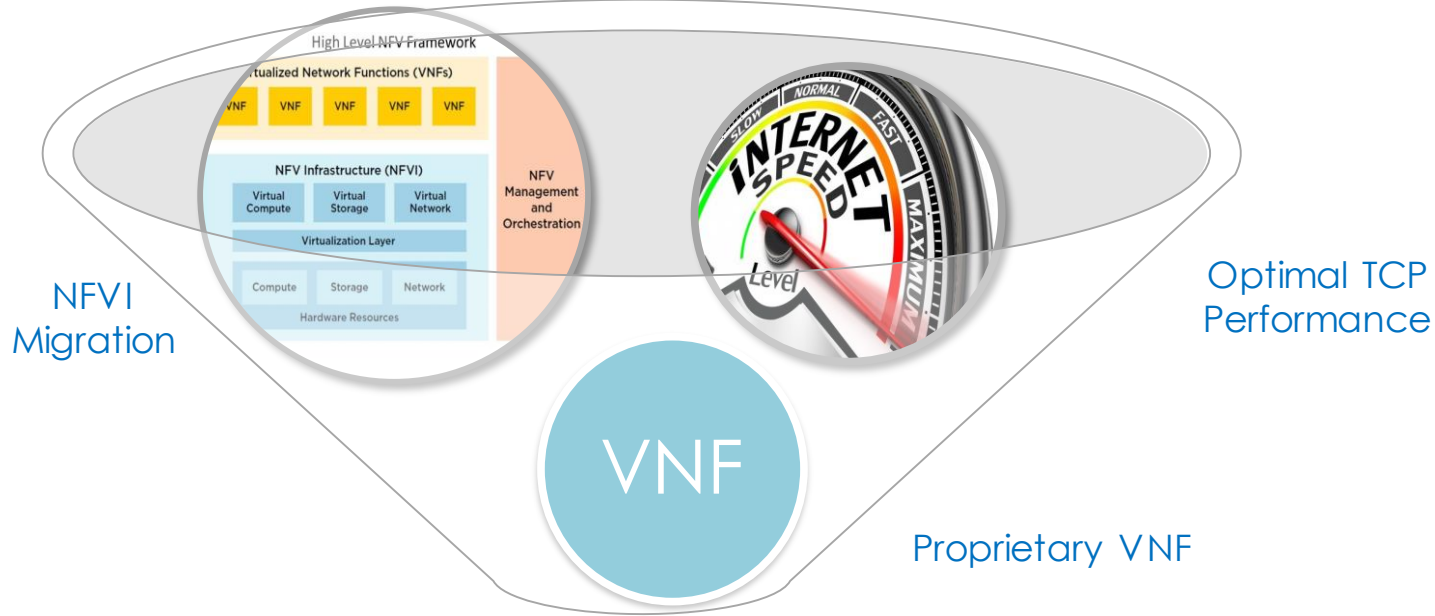
## **Technology Disclaimer:**

Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software or service activation. Performance varies depending on system configuration. No computer system can be absolutely secure. Check with your system manufacturer or retailer or learn more at [intel.com].

## **Performance Disclaimers:**

Cost reduction scenarios described are intended as examples of how a given Intel-based product, in the specified circumstances and configurations, may affect future costs and provide cost savings. Circumstances will vary. Intel does not guarantee any costs or cost reduction. Results have been estimated or simulated using internal Intel analysis or architecture simulation or modelling, and provided to you for informational purposes. Any differences in your system hardware, software or configuration may affect your actual performance.

# Problem Domain



# Deployment Scenario (Simplified)

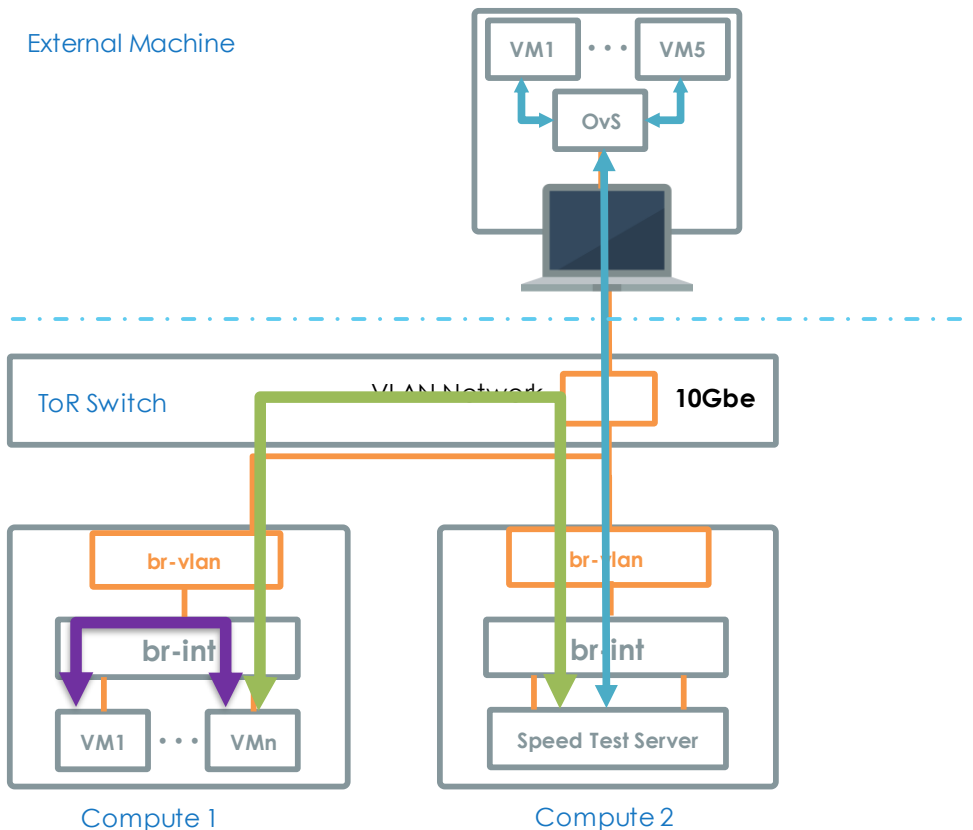
## Customer-Defined Test Cases

Speed Test Client → Speed Test Server

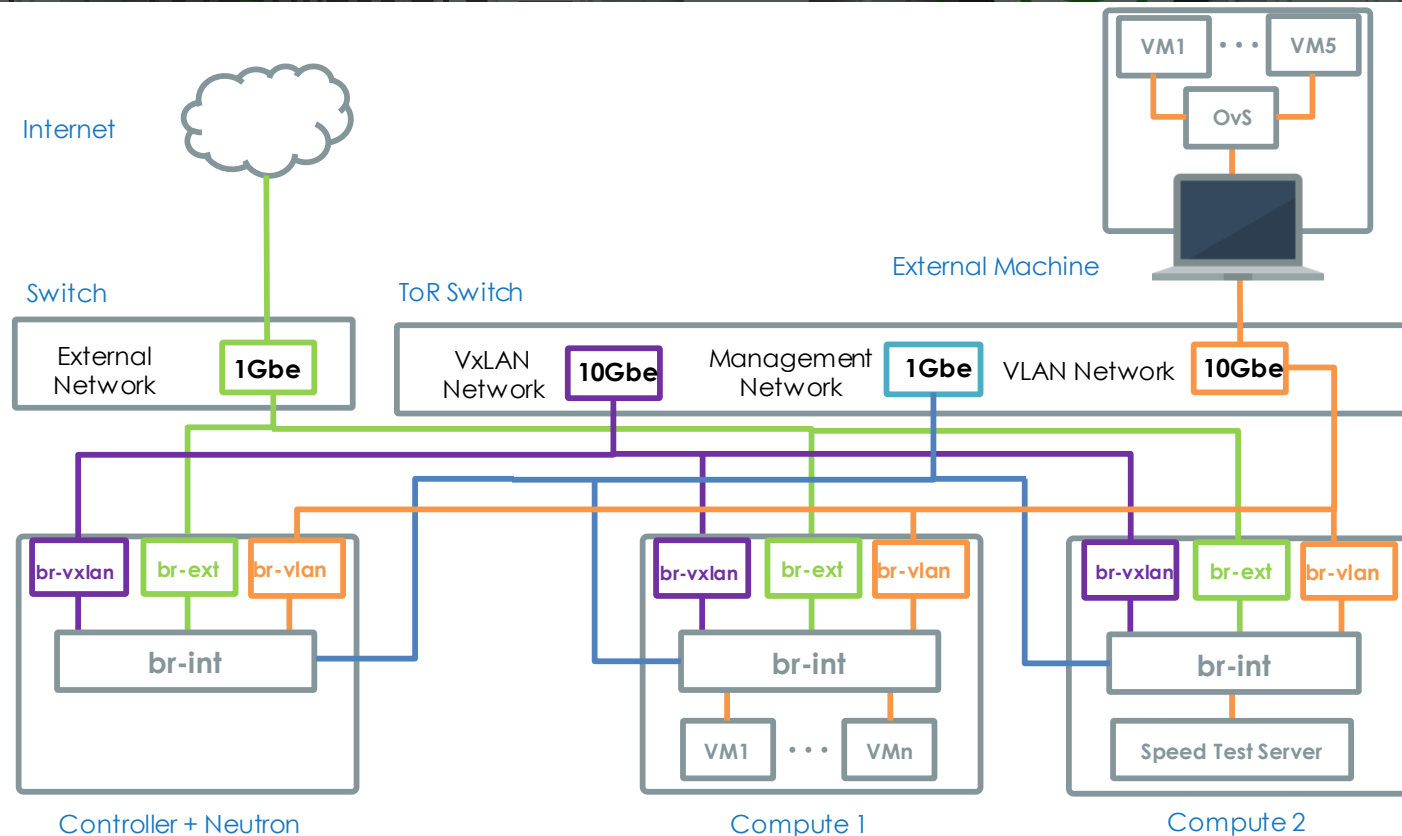
VM External Network → Speed Test Server Compute Node

Speed Test Server Compute Node 2 → VM Compute Node 1

VM Compute Node 1 → VM Compute Node 1

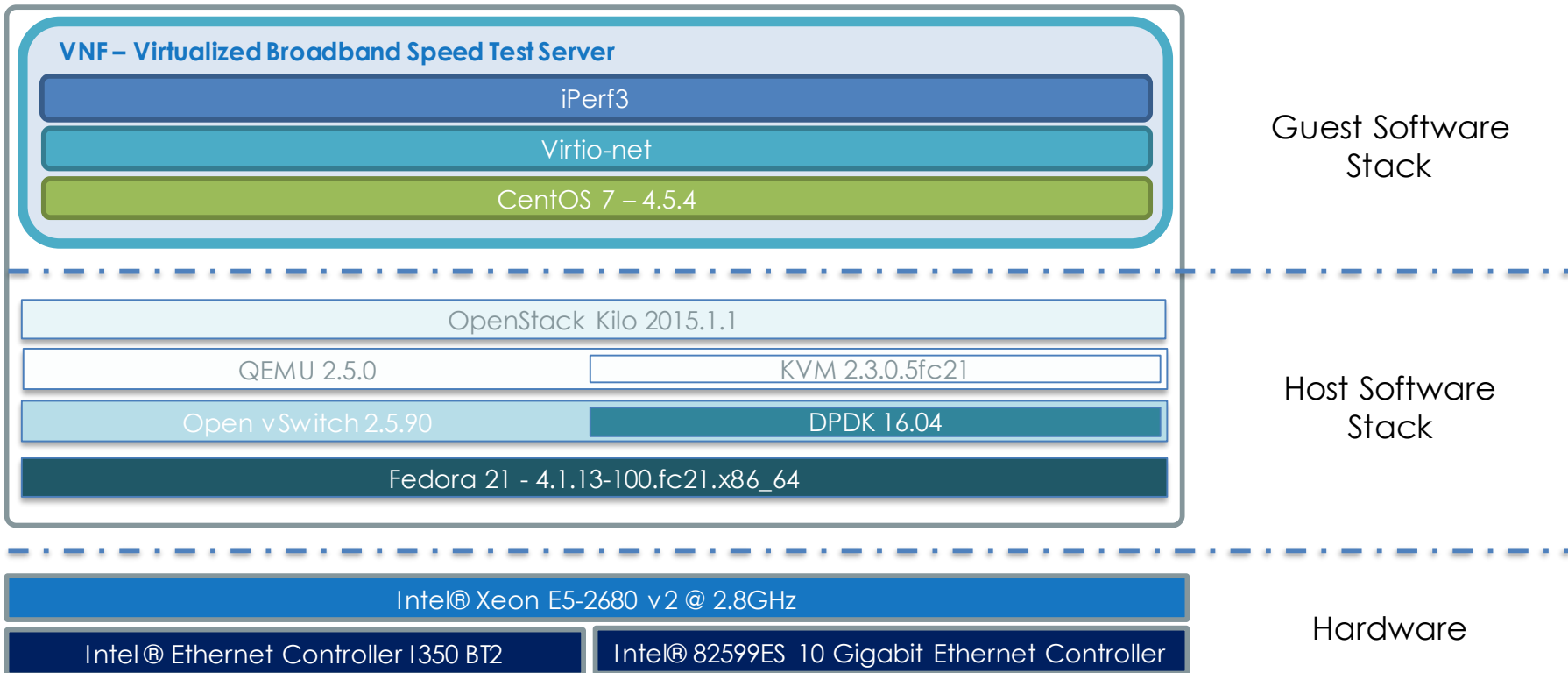


# VNF Deployment Scenario (Full)



# Anatomy of a VNF Compute Node

## Compute Node



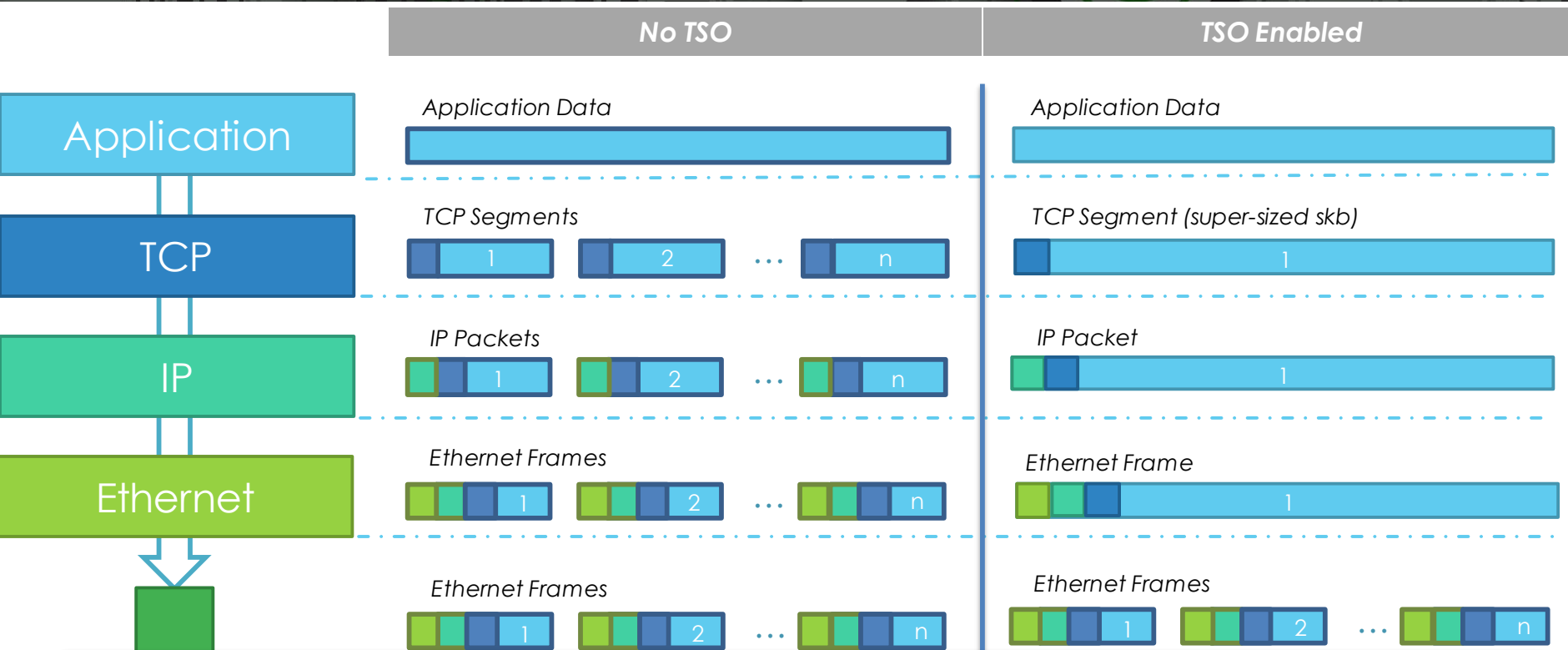
# Optimizations: Baseline

- ✓ Enable Hugepages
  - *Reduce the impact of Translation Lookaside Buffer (TLB) misses*
- ✓ Affinitize DPDK PMDs, and QEMU's virtual CPU threads
  - *Maximize CPU occupancy*
  - *Minimize cache thrashing*
- ✓ Enable NUMA support for OvS-DPDK
  - *Eliminate QPI traversal performance penalties*

Additional details available here 

<https://github.com/openvswitch/ovs/blob/master/INSTALL.DPDK-ADVANCED.md>

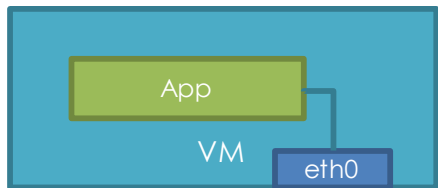
# Optimizations: TCP Segmentation Offload (TSO) Overview



Enable TSO in the guest to reduce vCPU load & boost throughput for OvS-DPDK



# Optimizations: TCP Segmentation Offload



```
ethtool -K eth0 tso on
```

Ethernet Frame



- ✓ Reduced vCPU load
- ✓ Improved PCI bus usage
- ✓ Higher throughput

GUEST



mbuf chain



HOST

```
mbuf->ol_flags & PKT_TX_TCP_SEG
```

```
mbuf->l2_len
```

```
mbuf->l3_len
```

```
mbuf->l4_len
```

```
mbuf->tso_segsz = MSS
```

```
mbuf->ol_flags |= PKT_TX_IP_CKSUM
```

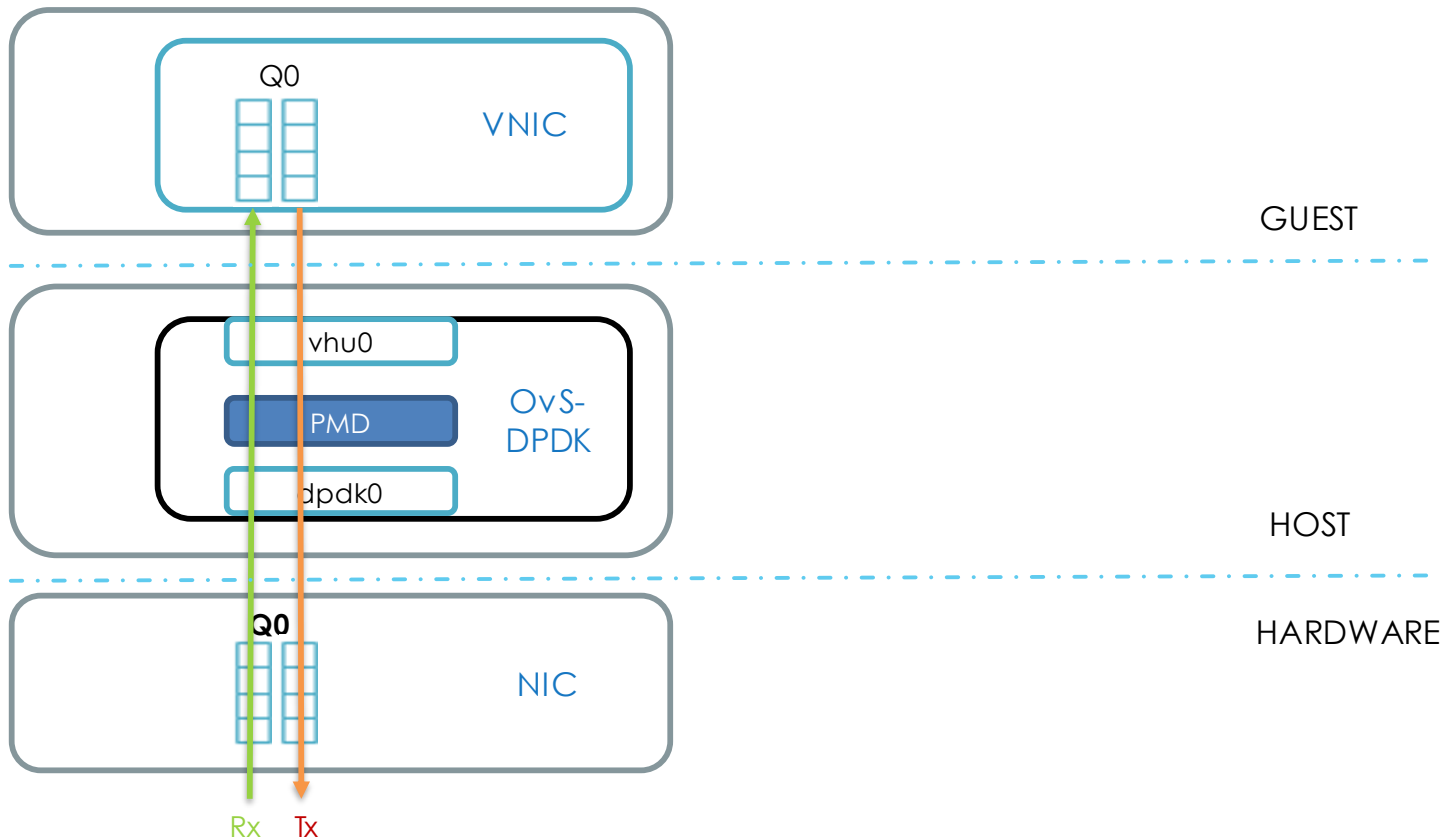
Ethernet Frames



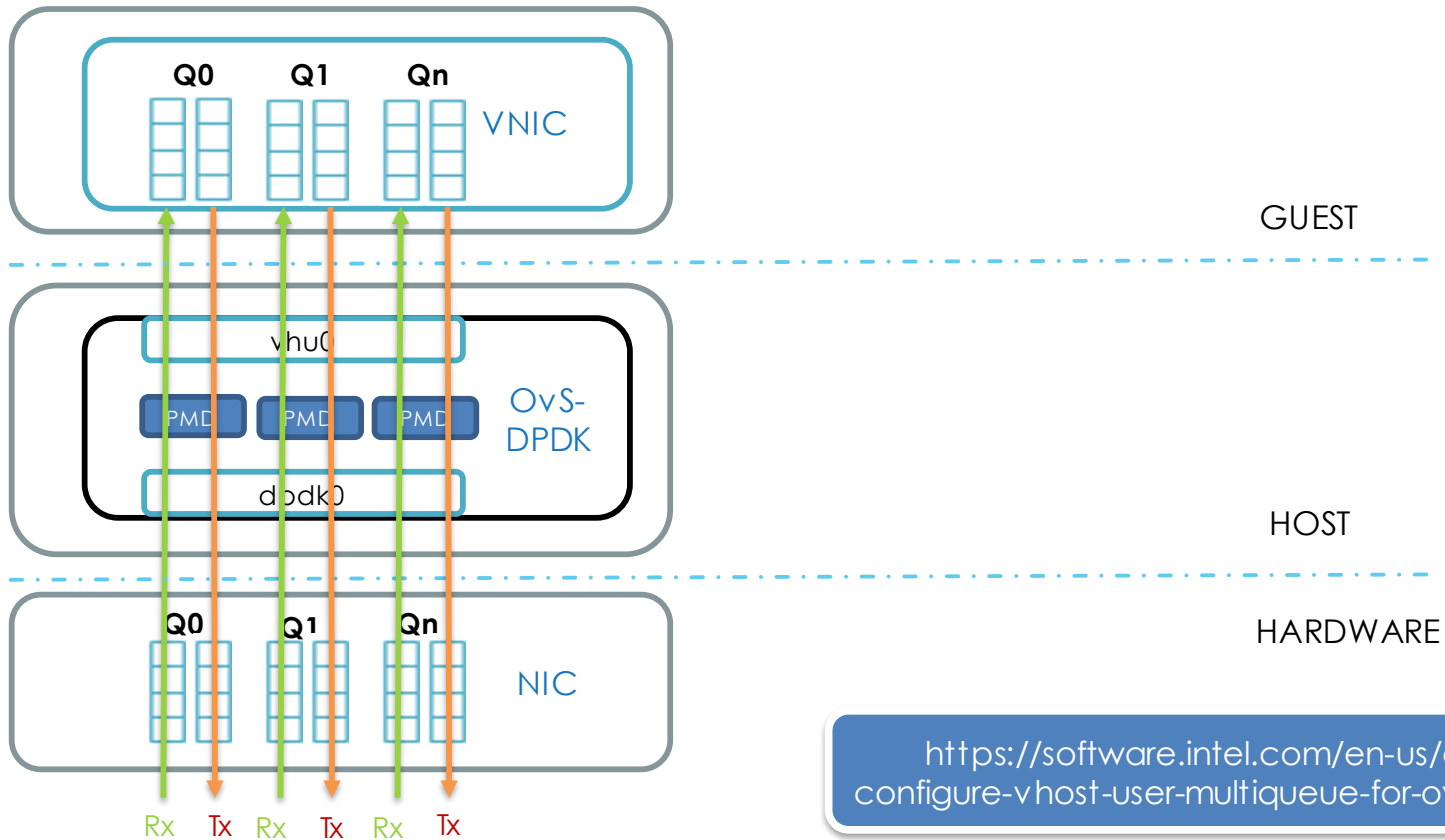
## RFC Patch

<https://mail.openvswitch.org/pipermail/ovs-dev/2016-June/235223.html>

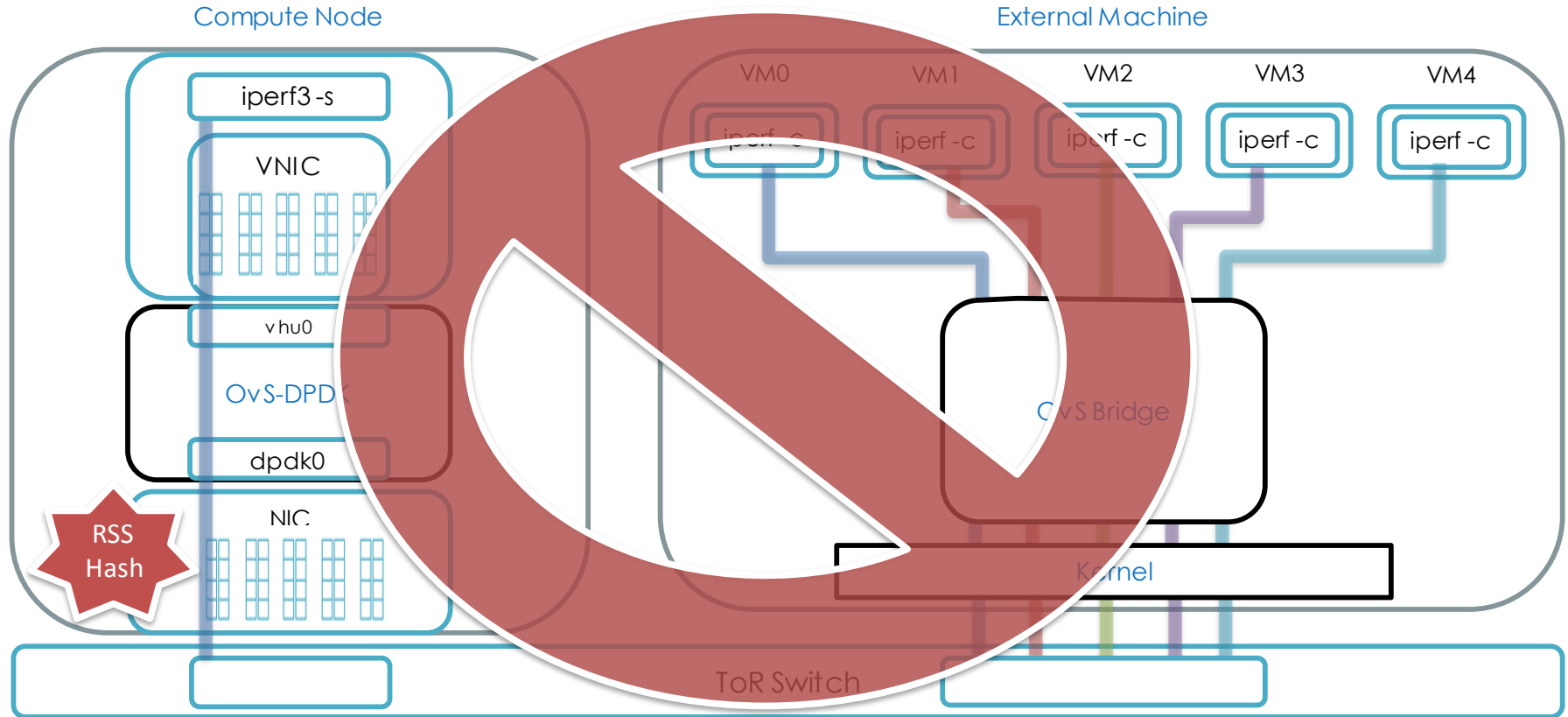
# TCP Optimizations: Multi Q (Overview)



# TCP Optimizations: Multi Q (Overview)



# TCP Optimizations: Multi Q (Problem)



# TCP Optimizations: Multi Q (Solution)

Compute Node

iperf3-s -P 10000

iperf3-s -P 10004

VNIC

vhu0

OvS-DPDK

dpgk0

RSS Hash

NIC

ToR Switch

External Machine

VM0

VM1

VM2

VM3

VM4

iperf -c  
-P 10000

iperf -c  
-P 10001

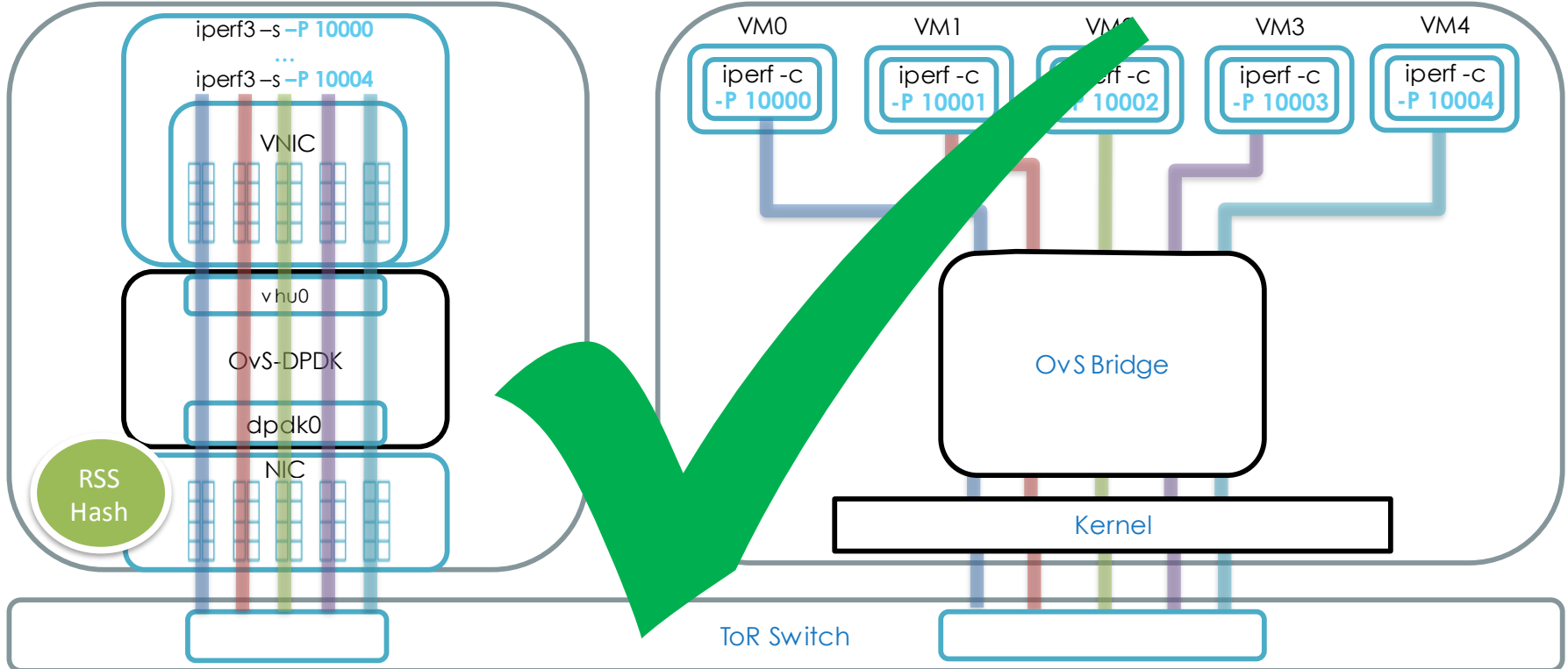
iperf -c  
-P 10002

iperf -c  
-P 10003

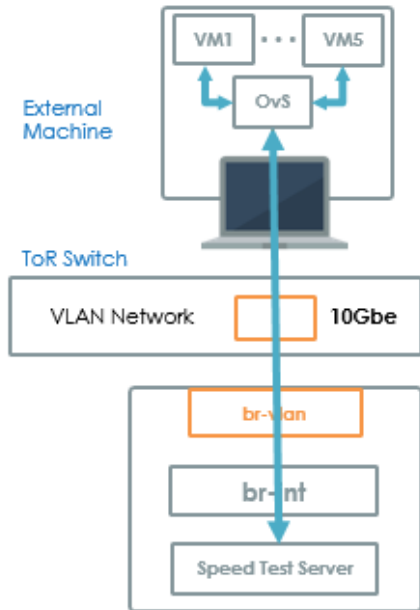
iperf -c  
-P 10004

OvS Bridge

Kernel



# Performance Results – Test Case #1



5 x EXTERNAL VM -> SINGLE SPEED TEST SERVER

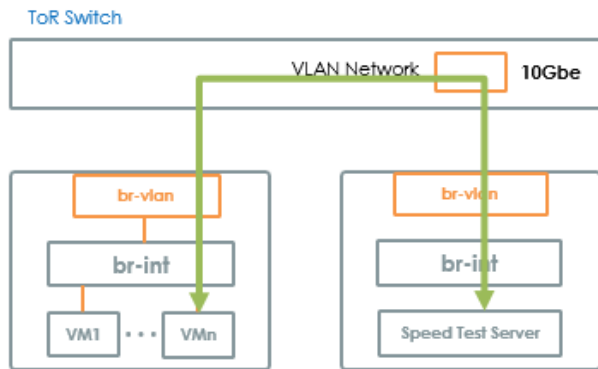
## AVERAGE SPEED TEST SERVER BANDWIDTH (GBPS)



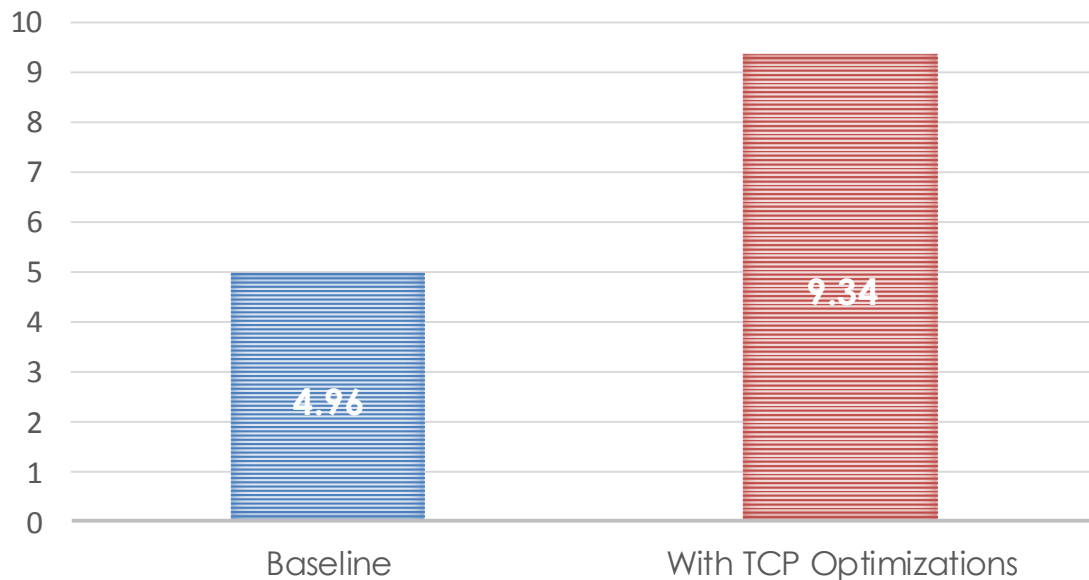
\*System configuration detailed in backup

# Performance Results – Test Case #2

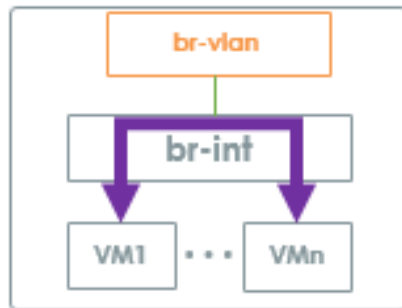
## AVERAGE SPEED TEST SERVER BANDWIDTH (GBPS)



SPEED TEST SERVER -> VM  
- SEPARATE COMPUTE NODES -

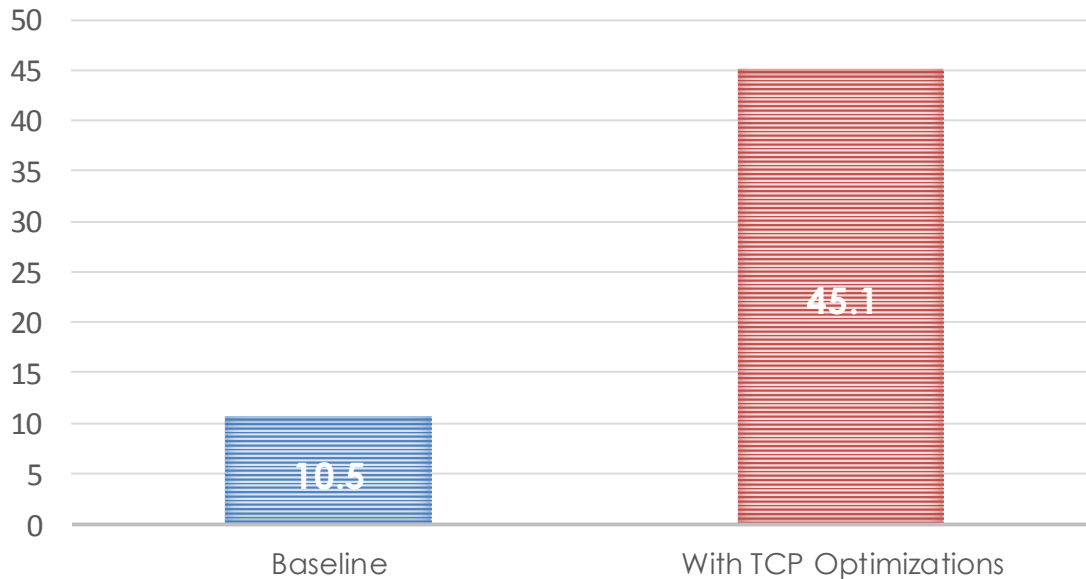


# Performance Results – Test Case #3



VM -> VM  
SAME COMPUTE NODE

## AVERAGE SPEED TEST SERVER BANDWIDTH (GBPS)





# Optimization Summary

## Baseline Optimizations

- Enable hugepages
- Per-port/RxQ PMD
- Affinitize workloads
- Incorporate NUMA support

## Avail of Offloads

- TSO = reduced vCPU load
- TSO = efficient PCI bandwidth consumption

## Utilize Multi Q for Guests

- Saturate line
- Push bottleneck back to the network

# Next Steps

- Release non-RFC TSO Support Patch
- Add support for TSO + Tunnels

# References

- <https://www.measurementlab.net/publications/understanding-broadband-speed-measurements.pdf>
- <https://software.intel.com/en-us/articles/configure-vhost-user-multiqueue-for-ovs-with-dpdk>
- <http://openvswitch.org/pipermail/dev/2016-June/072871.html>



↔ vs

Open vSwitch

Q&A



↔ vs

Open vSwitch

Backup

# System Configuration: Hardware

## Hardware Platform Specification

Server	Processor	Hard Drive	Memory	NIC
Compute1	Intel® Xeon® E5-2680 v2 at 2.80 GHz, 40 logical cores	1 TB	DDR3 1600 MHz	<ul style="list-style-type: none"><li>• Intel Ethernet Controller I350 BT2 (management and public networks)</li><li>• Intel® 82599 ES-10 Gigabit Ethernet Controller (VxLAN and VLAN networks)</li></ul>
Compute2	Intel® Xeon® E5-2680 v2 at 2.80 GHz, 40 logical cores	1 TB	DDR3 1600 MHz	<ul style="list-style-type: none"><li>• Intel Ethernet Controller I350 BT2 (management and public networks)</li><li>• Intel® 82599 ES-10 Gigabit Ethernet Controller (VxLAN and VLAN networks)</li></ul>

# System Configuration: Software

## Software Ingredients

#	Software BOM Item	Component
1	Operating System	Fedora* 21, Kernel 4.1.13-100.fc21.x86_64
2	Hypervisor	Compute nodes: QEMU-KVM, QEMU 2.5.0
3	Virtual Switch	Compute nodes: Open vSwitch 2.5.9+ <a href="#">TSO RFC patch</a>
4	Packet Processing Acceleration	DPDK v16.04
5	Virtualized Infrastructure Manager	OpenStack* Kilo 2015.1.0

# System Configuration: BIOS Settings 1/2

Apdio Setup Utility - Copyright (C) 2010 - 2013 America

**Processor Configuration**

Intel(R) QPI Link Frequency	8.0 GT/s
Intel(R) QPI Frequency Select	[Auto Max]
Intel(R) Turbo Boost Technology	[Enabled]
Enhanced Intel SpeedStep(R) Tech	[Enabled]
Processor C3	[Disabled]
Processor C6	[Enabled]
Intel(R) Hyper-Threading Tech	[Enabled]
Active Processor Cores	[All]
Execute Disable Bit	[Enabled]
Intel(R) Virtualization Technology	[Enabled]
Intel(R) VT for Directed I/O	[Enabled]
Interrupt Remapping	[Enabled]
Coherency Support	[Disabled]
ATS Support	[Enabled]
Pass-through DMA Support	[Enabled]
Intel(R) TXT	[Disabled]
Enhanced Error Containment Mode	[Disabled]
MLC Streamer	[Enabled]
MLC Spatial Prefetcher	[Enabled]
DCU Data Prefetcher	[Enabled]
DCU Instruction Prefetcher	[Enabled]
Direct Cache Access (DCA)	[Enabled]
Extended ATR	[0x03]
PFloor Tuning	12
SMM Wait Timeout	20

Apdio Setup Utility - Copyright (C) 2010 - 2013 America

**Power & Performance**

Power & Performance

CPU Power and Performance Policy [Balanced Performance]

[Performance] Optimization is strongly toward performance, even at the expense of energy efficiency.

[Balanced Performance] Weights optimization toward performance, while conserving energy.

[Balanced Power] Weights optimization toward energy conservation, with good performance.

[Power] Optimization is strongly toward energy efficiency, even at the expense of performance.

Apdio Setup Utility - Copyright (C) 2010 - 2013 America

**Memory Configuration**

Memory Configuration

Total Memory	64 GB
Effective Memory	65536 MB
Current Configuration	Independent
Current Memory Speed	DDR3-1600
Memory Operating Speed Selection	[Auto]
Phase Shedding	[Enabled]
Memory SPD Override	[Disabled]
Patrol Scrub	[Enabled]
Demand Scrub	[Enabled]
Correctable Error Threshold	[10]

► Memory RAS and Performance Configuration



# System Configuration: BIOS Settings 2/2

Aptio Setup Utility - Copyright (C) 2010 - 2013 American Megatronics  
Memory RAS and Performance Configuration

Memory RAS and Performance Configuration

Capabilities

Memory Mirroring Possible	YES
Memory Rank Sparing Possible	NO
Memory Lockstep Possible	YES
Select Memory RAS Configuration	[Maximum Performance]
NUMA Optimized	[Enabled]

Aptio Setup Utility - Copyright (C) 2010 - 2013 American Megatronics  
Socket 1 PCIe Ports Link Speed

Socket 1, DMI	[Gen 2 (5 GT/s)]
Socket 1, PCIe Port 1a	[Gen 3 (8 GT/s)]
Socket 1, PCIe Port 1b	[Gen 3 (8 GT/s)]
Socket 1, IO Module	[Gen 3 (8 GT/s)]
Socket 1, SAS Module	[Gen 3 (8 GT/s)]
Socket 1, PCIe Port 3a	[Gen 3 (8 GT/s)]
Socket 1, PCIe Port 3c	[Gen 3 (8 GT/s)]

Aptio Setup Utility - Copyright (C) 2010 - 2013 American Megatronics  
PCI Configuration

PCI Configuration

Maximize Memory below 4GB	[Disabled]
Memory Mapped I/O above 4GB	[Enabled]
Memory Mapped I/O Size	[Auto]
Onboard Video	[Enabled]
Legacy VGA Socket	[CPU Socket 1]
Dual Monitor Video	[Disabled]



↔ vs

Open vSwitch