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

- NFVI Migration
- Proprietary VNF
- Optimal TCP Performance
Deployment Scenario (Simplified)

Customer-Defined Test Cases

- **Speed Test Client** → **Speed Test Server**
- **VM** → **Speed Test Server** → **Compute Node**
- **Speed Test Server** → **VM**
- **VM** → **VM**

Diagram:
- **ToR Switch**
- **VLAN Network**
- **10Gbe**
- **VM1** → **VM5**
- **OvS**
- **External Machine**
- **br-vlan**
- **br-int**
- **Compute 1**
- **Compute 2**
VNF Deployment Scenario (Full)
Anatomy of a VNF Compute Node

**Hardware**

- **Intel® Xeon E5-2680 v2 @ 2.8GHz**
- **Intel® Ethernet Controller I350 BT2**
- **Intel® 82599ES 10 Gigabit Ethernet Controller**

**Compute Node**

- **VNF – Virtualized Broadband Speed Test Server**
  - iPerf3
  - Virtio-net
  - CentOS 7 – 4.5.4

**Host Software Stack**

- OpenStack Kilo 2015.1.1
- QEMU 2.5.0
- Open vSwitch 2.5.90
- Fedora 21 - 4.1.13-100.fc21.x86_64
- KVM 2.3.0.5fc21
- DPDK 16.04

**Guest Software Stack**

- iPerf3
- Virtio-net
- CentOS 7 – 4.5.4
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

<table>
<thead>
<tr>
<th>No TSO</th>
<th>TSO Enabled</th>
</tr>
</thead>
<tbody>
<tr>
<td>Application Data</td>
<td>Application Data</td>
</tr>
<tr>
<td>TCP Segments</td>
<td>TCP Segment (super-sized skb)</td>
</tr>
<tr>
<td>IP Packets</td>
<td>IP Packet</td>
</tr>
<tr>
<td>Ethernet Frames</td>
<td>Ethernet Frame</td>
</tr>
</tbody>
</table>

Enable TSO in the guest to reduce vCPU load & boost throughput for OvS-DPDK
Optimizations: TCP Segmentation Offload

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

**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
- vhu0
- OvS-DPDK
- dpdk0

NIC

RSS Hash

ToR Switch

External Machine

- VM0
  - iperf -c -P 10000
- VM1
  - iperf -c -P 10001
- VM2
  - iperf -c -P 10002
- VM3
  - iperf -c -P 10003
- VM4
  - iperf -c -P 10004

OvS Bridge

Kernel
Performance Results – Test Case #1

AVERAGE SPEED TEST SERVER
BANDWIDTH (GBPS)

- Client 1
- Client 2
- Client 3
- Client 4
- Client 5

Baseline
With TCP Optimizations

*System configuration detailed in backup
Performance Results – Test Case #2

AVERAGE SPEED TEST SERVER BANDWIDTH (GBPS)

Baseline: 4.96
With TCP Optimizations: 9.34

*System configuration detailed in backup
Performance Results – Test Case #3

AVERAGE SPEED TEST SERVER BANDWIDTH (GBPS)

VM -> VM
SAME COMPUTE NODE

Baseline: 10.5
With TCP Optimizations: 45.1

*System configuration detailed in backup
**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

## System Configuration: Hardware

### Hardware Platform Specification

<table>
<thead>
<tr>
<th>Server</th>
<th>Processor</th>
<th>Hard Drive</th>
<th>Memory</th>
<th>NIC</th>
</tr>
</thead>
</table>
| Compute1| Intel® Xeon® E5-2680 v2 at 2.80 GHz, 40 logical cores | 1 TB        | DDR3 1600 MHz  | • Intel Ethernet Controller I350 BT2 (management and public networks)  
                                          • Intel® 82599 ES-10 Gigabit Ethernet Controller (VxLAN and VLAN networks) |
| Compute2| Intel® Xeon® E5-2680 v2 at 2.80 GHz, 40 logical cores | 1 TB        | DDR3 1600 MHz  | • Intel Ethernet Controller I350 BT2 (management and public networks)  
                                          • Intel® 82599 ES-10 Gigabit Ethernet Controller (VxLAN and VLAN networks) |
## System Configuration: Software

### Software Ingredients

<table>
<thead>
<tr>
<th>#</th>
<th><strong>Software BOM Item</strong></th>
<th><strong>Component</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Operating System</td>
<td>Fedora* 21, Kernel 4.1.13-100.fc21.x86_64</td>
</tr>
<tr>
<td>2</td>
<td>Hypervisor</td>
<td>Compute nodes: QEMU-KVM, QEMU 2.5.0</td>
</tr>
<tr>
<td>3</td>
<td>Virtual Switch</td>
<td>Compute nodes: Open vSwitch 2.5.9+ [TSO RFC patch]</td>
</tr>
<tr>
<td>4</td>
<td>Packet Processing Acceleration</td>
<td>DPDK v16.04</td>
</tr>
<tr>
<td>5</td>
<td>Virtualized Infrastructure Manager</td>
<td>OpenStack* Kilo 2015.1.0</td>
</tr>
</tbody>
</table>
### System Configuration: BIOS Settings 1/2

#### Processor Configuration

<table>
<thead>
<tr>
<th>Feature</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intel(R) QPI Link Frequency</td>
<td>8.0 GT/s</td>
</tr>
<tr>
<td>Intel(R) QPI Frequency Select</td>
<td>[Auto Max]</td>
</tr>
<tr>
<td>Intel(R) Turbo Boost Technology</td>
<td>[Enabled]</td>
</tr>
<tr>
<td>Enhanced Intel SpeedStep(R) Tech</td>
<td>[Enabled]</td>
</tr>
<tr>
<td>Processor C3</td>
<td>[Disabled]</td>
</tr>
<tr>
<td>Processor C6</td>
<td>[Enabled]</td>
</tr>
<tr>
<td>Intel(R) Hyper-Threading Tech</td>
<td>[Enabled]</td>
</tr>
<tr>
<td>Active Processor Cores</td>
<td>[All]</td>
</tr>
<tr>
<td>Execute Disable Bit</td>
<td>[Enabled]</td>
</tr>
<tr>
<td>Intel(R) Virtualization Technology</td>
<td>[Enabled]</td>
</tr>
<tr>
<td>Intel(R) VT for Directed I/O</td>
<td>[Enabled]</td>
</tr>
<tr>
<td>Interrupt Remapping</td>
<td>[Enabled]</td>
</tr>
<tr>
<td>Coherency Support</td>
<td>[Disabled]</td>
</tr>
<tr>
<td>ATS Support</td>
<td>[Enabled]</td>
</tr>
<tr>
<td>Pass-through DMA Support</td>
<td>[Enabled]</td>
</tr>
<tr>
<td>Intel(R) TXT</td>
<td>[Disabled]</td>
</tr>
<tr>
<td>Enhanced Error Containment Mode</td>
<td>[Disabled]</td>
</tr>
<tr>
<td>LLC Streamer</td>
<td>[Enabled]</td>
</tr>
<tr>
<td>MLC Spatial Prefetcher</td>
<td>[Enabled]</td>
</tr>
<tr>
<td>DCU Data Prefetcher</td>
<td>[Enabled]</td>
</tr>
<tr>
<td>DCU Instruction Prefetcher</td>
<td>[Enabled]</td>
</tr>
<tr>
<td>Direct Cache Access (DCA)</td>
<td>[Enabled]</td>
</tr>
<tr>
<td>Extended ATR</td>
<td>[16x]</td>
</tr>
<tr>
<td>PFloor Tuning</td>
<td>[20]</td>
</tr>
<tr>
<td>SMM Wait Timeout</td>
<td>[20]</td>
</tr>
</tbody>
</table>

#### Power & Performance

- **CPU Power and Performance Policy**: [Balanced Performance]
- **Performance Optimization**: is strongly towards performance, even at the expense of energy efficiency.
- **Balanced Performance**: Weights optimization towards performance, while conserving energy.
- **Balanced Power**: Weights optimization towards energy conservation, with good performance.
- **Power Optimization**: is strongly towards energy efficiency, even at the expense of performance.

#### Memory Configuration

<table>
<thead>
<tr>
<th>Feature</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Memory</td>
<td>64 GB</td>
</tr>
<tr>
<td>Effective Memory</td>
<td>65536 MB</td>
</tr>
<tr>
<td>Current Configuration</td>
<td>Independent</td>
</tr>
<tr>
<td>Current Memory Speed</td>
<td>DDR3-1600</td>
</tr>
<tr>
<td>Memory Operating Speed Selection</td>
<td>[Auto]</td>
</tr>
<tr>
<td>Phase Shading</td>
<td>[Enabled]</td>
</tr>
<tr>
<td>Memory SPD Override</td>
<td>[Disabled]</td>
</tr>
<tr>
<td>Patrol Scrub</td>
<td>[Enabled]</td>
</tr>
<tr>
<td>Demand Scrub</td>
<td>[Enabled]</td>
</tr>
<tr>
<td>Correctable Error Threshold</td>
<td>[10]</td>
</tr>
</tbody>
</table>

#### Memory RAS and Performance Configuration
System Configuration: BIOS Settings 2/2

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]

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)]

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]