NSX Performance

Samuel Kommu
Disclaimer

- This presentation may contain product features that are currently under development.
- This overview of new technology represents no commitment from VMware to deliver these features in any generally available product.
- Features are subject to change, and must not be included in contracts, purchase orders, or sales agreements of any kind.
- Technical feasibility and market demand will affect final delivery.
- Pricing and packaging for any new technologies or features discussed or presented have not been determined.
Objectives

1. NSX Performance Characteristics
2. Recommendations for Optimal Performance
3. Benchmarking Considerations
## Agenda

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>NSX – Overview of the Overlays</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>Performance Tuning</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>VMware NIC Compatibility Guide</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>East-West Performance</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>North-South Performance</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>IETF</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Summary &amp; QA</td>
<td></td>
</tr>
</tbody>
</table>
VXLAN Frame Format

VXLAN Encapsulated Frame

Original Ethernet Frame

IP over UDP encapsulation

VTEPs*
Geneve Frame Format

Geneve Encapsulated Frame

IP over UDP encapsulation
Performance Tuning
Parameters that matter -
MTU / MSS
Maximum Transmission Unit / Maximum Segment Size

ESX 1
- Virtual Machine
  - vNIC
  - MTU: 1600/9000
  - 40G pNIC

ESX 2
- Virtual Machine
  - vNIC
  - MTU: 1600/9000
  - 40G pNIC

40G Switch
Maximum Packet Size Allowed
**TSO for Overlay Traffic**

**NIC Based TSO**

- MAC
- IP
- UDP
- VXLAN
- MAC
- IP
- TCP
- Payload

**CPU Based TSO**

- MAC
- IP
- UDP
- VXLAN
- MAC
- IP
- TCP
- Payload

**Physical Fabric**

For best performance – use Hardware TSO for Overlay Traffic
LRO for Overlay Traffic

NIC Based LRO

VM

Software LRO available on ESX 6.5

All components in this path including Logical Switch, Firewall etc., work with 32K segments

QLogic 45000 Series
Support LRO in Hardware

Physical Fabric

QLogic 45000 Series Support LRO in Hardware

All components in this path including Logical Switch, Firewall etc., work with 32K segments

VMworld 2017 Content: Not for publication or distribution
Receive Side Scaling (RSS)

- Without Receive side scaling
  - All traffic is handled by a single core
  - Multiple cores are not used even if available
  - Traffic throughput slows down based on a single core’s capacity
Receive Side Scaling (RSS)

- With Receive Side Scaling Enabled
  - Network adapter has multiple queues to handle receive traffic
  - 5 tuple based hash (Src/Dest IP, Src/Dest MAC and Src Port) for optimal distribution to queues
  - Kernel thread per receive queue helps leverage multiple CPU cores
Rx/Tx Filters

- Uses inner packet headers to queue traffic
Native Driver

VMKernel Data Structures

Translation

VMKLinux Driver

Physical Fabric

MAC IP UDP VXLAN MAC IP TCP Payload

1500/9000

Less CPU cycles spent on translating

With ESX 6.5 Native Driver
# vNICs vs # of Queues

![Diagram](image)

Scale up throughput by using multiple vNICs

ESX 1

pNIC

vNIC 1

vNIC 2

VM
# of Queues on a single vNIC

Scale up throughput on a single vNIC by leveraging multiple Queues
VMware NIC Compatibility Guide
What features does my NIC card support?
How To – VMware Compatibility Guide

Where is it?

How To – VMware Compatibility Guide

Intel 710: Search for a driver supporting VXLAN-Offload

Select the Version of ESX
Select brand name
Card model name – if available
Click on Update and View Results

Features of interest
I/O Device Type

Click on Update and View Results
How To – VMware Compatibility Guide
Intel 710: Search results for driver supporting VXLAN-Offload

<table>
<thead>
<tr>
<th>Intel</th>
<th>Intel(R) Ethernet Controller X710 for 10GbE backplane</th>
<th>Network</th>
<th>ESX</th>
<th>ESXi</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>U2</td>
<td>U1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>4.0</td>
<td>6.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3.0.2</td>
<td>6.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3.0.1</td>
<td>6.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3.0</td>
<td>5.5</td>
</tr>
</tbody>
</table>

From the results – Select the specific Card – 10GbE SFP+

<table>
<thead>
<tr>
<th>Intel</th>
<th>Intel(R) Ethernet Controller X710 for 10GbE SFP+</th>
<th>Network</th>
<th>ESX</th>
<th>ESXi</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>U2</td>
<td>U1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>4.0</td>
<td>6.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3.0.2</td>
<td>6.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3.0.1</td>
<td>6.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3.0</td>
<td>5.5</td>
</tr>
</tbody>
</table>

Click on the ESX version

<table>
<thead>
<tr>
<th>Intel</th>
<th>Intel(R) Ethernet Controller XL710 for 10GbE QSFP+</th>
<th>Network</th>
<th>ESX</th>
<th>ESXi</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>U2</td>
<td>U1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>4.0</td>
<td>6.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3.0.2</td>
<td>6.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3.0.1</td>
<td>6.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3.0</td>
<td>5.5</td>
</tr>
</tbody>
</table>

Click on the ESX version
How To – VMware Compatibility Guide

Intel 710: Search results for driver supporting VXLAN-Offload - Continued

Click on [+] to expand and check the features supported.
How To – VMware Compatibility Guide

Intel 710: Search results for driver supporting VXLAN-Offload – Found!

<table>
<thead>
<tr>
<th>Model Release Details</th>
<th>Expand All</th>
<th>Collapse All</th>
</tr>
</thead>
<tbody>
<tr>
<td>VMware Product Name</td>
<td>ESXi 6.0 U2</td>
<td></td>
</tr>
<tr>
<td>Release</td>
<td>Device Driver(s)</td>
<td>Firmware Version</td>
</tr>
<tr>
<td>ESXi 6.0 U2</td>
<td>i40e version 1.4.28</td>
<td>5.04</td>
</tr>
</tbody>
</table>


Looks like this one supports a bunch of features ... Yay!!
Traffic Profiles
In a Datacenter
End-to-End Traffic Flows

East – West & North - South

Mix of bandwidth hungry and other Flows

ESX 1

LS1-VM1 → LS1-VM2 → LS1-VM8

ESX 2

LS2-VM1 → LS2-VM2 → LS2-VM8

NSX EDGE

Larger Mix of Packets
Higher PPS Requirements

pNIC

ToR

Rest of DC/Cloud
Typical Applications and Traffic Profiles

In Datacenter

• Long Flows – Designed to maximize on bandwidth
  – Logs
  – Backups
  – FTP
  – Web Servers

• Short Flows – Lower bandwidth requirements
  – Databases
    • Specially in memory ones – or cache layers –
    • even in those cases bulk requests are made for efficiency – refer FB on Memcache

• Small Packets (<200 Bytes) – Generally no SLAs – Some rare cases where latency is important
  – DNS
  – DHCP
  – TCP ACKs
  – Keep Alive Messages

Mix of different size packets
Larger throughput heavy flows
Smaller Latency Sensitive Flows

Majority of the packets – at this packet size
## Typical Applications and Traffic Profiles

### Physical vs Virtual

<table>
<thead>
<tr>
<th></th>
<th>Physical Fabric (Any)</th>
<th>Virtual Fabric (E/W)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Long Flows</strong></td>
<td>~ 1500 Bytes</td>
<td>32K – 64K (TSO, LRO)</td>
</tr>
<tr>
<td>(&gt;1500 Bytes)</td>
<td></td>
<td>By Default</td>
</tr>
<tr>
<td></td>
<td>Throughput Focused</td>
<td>May be tuned to even higher values</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(NICs do the heavy lifting)</td>
</tr>
<tr>
<td><strong>Short Flows</strong></td>
<td>Packet Size</td>
<td>Packet Size</td>
</tr>
<tr>
<td>(&lt;=1500 Bytes)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Not Bandwidth Hungry</td>
<td></td>
</tr>
</tbody>
</table>

*Not for publication or distribution*
# Application Layer Performance Benchmarking Tools

Which ones and why

<table>
<thead>
<tr>
<th>*</th>
<th>Relevance for Virtual Infra (E/W)</th>
</tr>
</thead>
<tbody>
<tr>
<td>iPerf 2</td>
<td>Benchmarks: TSO/LRO Implementation, NIC Cards, Multiple Cores</td>
</tr>
<tr>
<td>iPerf 3</td>
<td>Similar to iPerf2 – Single core Benchmarking only</td>
</tr>
<tr>
<td>NetPerf</td>
<td>Similar to iPerf – with a few more bells and whistles</td>
</tr>
<tr>
<td>PktGen</td>
<td>For some of the N/S Benchmarking – Specially with BM Edge</td>
</tr>
<tr>
<td>Application Level</td>
<td>Application level benchmarks for throughput and latency</td>
</tr>
</tbody>
</table>

- Apache Benchmark
- Memcache
East – West
Performance Characteristics
## Setup for East-West Throughput Performance

<table>
<thead>
<tr>
<th>Servers</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>OS</td>
<td>ESXi 6.5 U1</td>
</tr>
<tr>
<td>Processor</td>
<td>Intel(R) Xeon(R) CPU E5-2699 v4 @ 2.20GHz</td>
</tr>
<tr>
<td>Hyper-Threading</td>
<td>Enabled</td>
</tr>
<tr>
<td>RAM</td>
<td>256 GB</td>
</tr>
<tr>
<td>MTU</td>
<td>9000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Virtual Machines</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>OS</td>
<td>RedHat 6</td>
</tr>
<tr>
<td>vCPU</td>
<td>2</td>
</tr>
<tr>
<td>RAM</td>
<td>2</td>
</tr>
<tr>
<td>Network</td>
<td>VMXNET3</td>
</tr>
<tr>
<td>MTU</td>
<td>1500 / 8900</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>NIC Card Details</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Intel</td>
<td>XL710 (40 GbE)</td>
</tr>
<tr>
<td>Driver</td>
<td>I40e</td>
</tr>
<tr>
<td>Version</td>
<td>1.4.3</td>
</tr>
</tbody>
</table>
End-to-End Traffic Flows

East – West Traffic Flow

LS1-VM1 → LS1-VM2 → LS1-VM8

ESX 1

LS2-VM1 → LS2-VM2 → LS2-VM8

ESX 2

pNIC

Switch

Rest of DC/Cloud

Larger Mix of Packets
Higher PPS Requirements

pNIC

NSX EDGE
Topology – Logical Switch
Topology – Tier 1
Topology – Tier 0

- Host 1
  - VM1
  - VM2
  - .......
  - VM4
- Tier-0 Router
- Logical Switch 1
- Logical Switch 2
- Host 2
  - VM1
  - VM2
  - .......
  - VM4
TCP Throughput – ESX (East - West)

Benchmarking Methodology:
• iPerf 2.0.5
  • Options “-P 4 -t 30”
  • Across 4 VM Pairs
  • 4 Threads per VM Pair

Observations
• ~25Gbps with 1500 MTU
• Close to line rate with 9K MTU

To achieve ~80Gbps
• Use 2 x 40GbE NIC cards
North - South
VM Edge
Edge is DPDK enabled

- High forwarding performance
  - Pull mode driver,
  - queue manager,
  - buffer manager, etc.)
- Linear performance increase by addition of cores
- More info can be found at
  - http://www.intel.com/go/DPDK
Fast Path

Without a Hash Table
For cluster of packets that arrive together

With a Hash Table
For an entire Flow
ESG Performance Environment

- Simple end to end topology
  - Single Edge Gateway
  - Two Uplinks
    - One for overlay and other for VLAN

- Performance Measurement
  - iPerf
  - True application TCP performance
  - Multi-thread, multi-VM

- Single Edge Gateway Characteristics
  - Hyper Threading Disabled
  - CPU power saving disabled
  - 4 vCPU Form Factor

- VM Config
  - MTU change based on benchmark
  - VMXNET 3

<table>
<thead>
<tr>
<th>Servers</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>ESX</td>
<td>6.0. U2 / 6.5 U1</td>
</tr>
<tr>
<td>NSX</td>
<td>6.2 / NSX-T</td>
</tr>
<tr>
<td>Server Make/Model</td>
<td>Dell PowerEdge R720 / Intel</td>
</tr>
<tr>
<td>Processor</td>
<td>2 x Intel ® Xeon ® CPU E5-2680 v2 @ 2.80GHz / 2 x E5 2699 v4 @ 2.20Ghz</td>
</tr>
<tr>
<td>Hyper-Threading</td>
<td>Enabled (Disabled on ESG Node)</td>
</tr>
<tr>
<td>RAM</td>
<td>128 GB</td>
</tr>
<tr>
<td>MTU</td>
<td>1600 or 9000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>NIC Card Details</th>
<th>Virtual Machines</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intel XL710</td>
<td>OS</td>
</tr>
<tr>
<td>Driver i40e</td>
<td>RHEL 6 (64-bit)</td>
</tr>
<tr>
<td>Version 1.4.26-1OEM.550.0.0.1331820</td>
<td>vCPU 2</td>
</tr>
<tr>
<td>MTU 1600 or 9000</td>
<td>RAM 2 GB</td>
</tr>
<tr>
<td></td>
<td>Network VMXNET3</td>
</tr>
<tr>
<td></td>
<td>MTU 1500</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ESG Details</th>
<th>iPerf</th>
</tr>
</thead>
<tbody>
<tr>
<td>Form Factor</td>
<td>Version 2.0.5</td>
</tr>
<tr>
<td>Quad-Large / Medium (4 vCPU)</td>
<td></td>
</tr>
</tbody>
</table>
### Topology

**NSX-V Edge**

- **VM3** (172.16.214.90)
- **VM4** (172.16.214.91)
- **VM1** (20.20.20.3)
- **VM2** (20.20.20.6)

**ESX 1**

**40G pNIC**

**40G Switch**

**ESX 2**

**40G pNIC**

**VLAN**

**VXLAN**

**VDS**

**ESG**

**DLR**

**VPN**

**ESG**

**VXLAN**

**VLAN**
Topology

NSX-T VM Edge
Edge Throughput

- ESG is not limited to 10G
- Can take advantage of 25/40G Ports
- 9K MTU for close to line rate throughput
- End to End tuning for best results
  - Hard Drive Speed
  - Application tuning

Setup Details:
- 4 x vCPU (Quad Large / Medium)
  - Hyper-threading Disabled on host
- iPerf 2
- 2 x Intel® E5-2680 v2 @ 2.80GHz (NSX-v)
- 2 x Intel® E5-2699 v4 @ 2.20GHz (NSX-t)
- Sender VMs / Receiver VMs
  - 2 x vCPU per VM
- Intel XL710 40G Card
North – South
Bare Metal Edge
## Setup for North-South Throughput Performance

<table>
<thead>
<tr>
<th>Servers</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>OS</td>
<td>Bare Metal Edge</td>
</tr>
<tr>
<td>Processor</td>
<td>Intel(R) Xeon(R) CPU E5-2699 v4 @ 2.20GHz</td>
</tr>
<tr>
<td>Hyper-Threading</td>
<td>Enabled</td>
</tr>
<tr>
<td>RAM</td>
<td>256 GB</td>
</tr>
<tr>
<td>MTU</td>
<td>1500</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Virtual Machines</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>OS</td>
<td>RedHat 6</td>
</tr>
<tr>
<td>vCPU</td>
<td>2</td>
</tr>
<tr>
<td>RAM</td>
<td>2</td>
</tr>
<tr>
<td>Network</td>
<td>VMXNET3</td>
</tr>
<tr>
<td>MTU</td>
<td>1500 / 8900</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>NIC Card Details</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Intel</td>
<td>x520</td>
</tr>
<tr>
<td>Driver</td>
<td>ixgbe</td>
</tr>
<tr>
<td>Version</td>
<td>In-Box</td>
</tr>
</tbody>
</table>
Bare Metal Edge

Topology

Bandwidth Usage
Measured Here

MTU 1500

VMworld 2017 Content: Not for publication or distribution
Throughput

256 Bytes

Line rate @ 10 Gbps Port

Not the true limit -
Bottleneck at sender/receiver side -
## Setup for North-South Throughput Performance

<table>
<thead>
<tr>
<th>Servers</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Processor</td>
<td>Intel ® Xeon ® CPU E5-2667 v2 @ 3.30GHz</td>
</tr>
<tr>
<td># of Sockets</td>
<td>2</td>
</tr>
<tr>
<td># of Logical Cores</td>
<td>16</td>
</tr>
<tr>
<td>Hyper-Threading</td>
<td>Disabled</td>
</tr>
<tr>
<td>RAM</td>
<td>64 GB</td>
</tr>
<tr>
<td>NIC</td>
<td>Intel X540-AT2</td>
</tr>
</tbody>
</table>
IETF
Benchmarking Considerations for Virtual Infrastructures
Internet Engineering Task Force – Active Internet-Draft

Please share your comments …
In Summary ...
In Summary
Inter Host Bottlenecks ...

- Single Core Limits
  - 5 to 20Gbps per core based on MTU
- Multi Core
  - 4 X times of single core limits
  - Can go slightly beyond 80G
- PCIe 3.0 Limitations
  - ~8Gbps per lane
  - Most NICs are x8 lanes
    - ~64 Gbps limit
    - Use two NICs for > 40G Throughput
- For high throughput
  - 2 x 8 lane NICs or 1 x 16 lane NIC
  - Higher MTU
  - TSO, LRO & RSS enabled cards
Summary

• Closer to the Application Layer
  – Large packets processed at source with very little overhead

• It's not all software
  – Offloads
    • RSS
    • VXLAN
    • TSO
    • LRO
    • SSL

• Hardware and Application Limitations
  – Use 2 PCIe Slots instead of 1 – if looking for greater than 40G bandwidth
  – Figure out what’s really important for your Application – Throughput? Latency?
Please fill out your survey.

Take a survey and enter a drawing for a VMware company store gift card.
Thank You