Considerations for Benchmarking Virtual Networks

draft-bmwg-nvp-03

Samuel Kommu, skommu@vmware.com

Jacob Rapp, jrapp@vmware.com

July 2018 IETF 102 – Montreal

Considerations for Benchmarking Network Virtualization Platforms - Overview

draft-bmwg-nvp-03

Why : Physical vs Virtual Network Platforms - Differences

MTU limited packets vs Higher Level Segments

Scope

Hypervisor Based Network Virtualization Platforms only – Not NFV

Considerations

Application Layer Benchmarks

Working closer to application layer segments and not low level packets

Server Hardware

Support for HW offloads (TSO / LRO / RSS)

Other Hardware offload benefits – Performance Related Tuning

Frame format sizes within Hypervisor

Scale Testing for New Application Architectures

New micro-Service type architectures

Documentation

System Under Test vs Device Under Test

Intra-Host (Source and destination on the same host)

Inter-Host (Source and Destination on different hosts – Physical Infra providing connectivity is part of SUT)

Changes from previous draft

draft-bmwg-nvp-03

Scope

Most of comments and questions were around clarifying scope

These benchmark considerations are specific to two scenarios of Network Virtualization Edge (NVE)

1. NVE Co-located with the server hypervisor (RFC 8014 Section 4.1 **An Architecture for Data-Center Network Virtualization over Layer 3 (NVO3)**) – "When server virtualization is used, the entire NVE functionality will typically be implemented as part of the hypervisor and/or virtual switch on the server. "

2. Split-NVE (RFC 8394 **Split Network Virtualization Edge (Split-NVE) Control-Plane Requirements** Section 1.1) – "Another possible scenario leads to the need for a split-NVE implementation. An NVE running on a server (e.g., within a hypervisor) could support NVO3 service towards the tenant but not perform all NVE functions (e.g., encapsulation) directly on the server; some of the actual NVO3 functionality could be implemented on (i.e., offloaded to) an adjacent switch to which the server is attached."

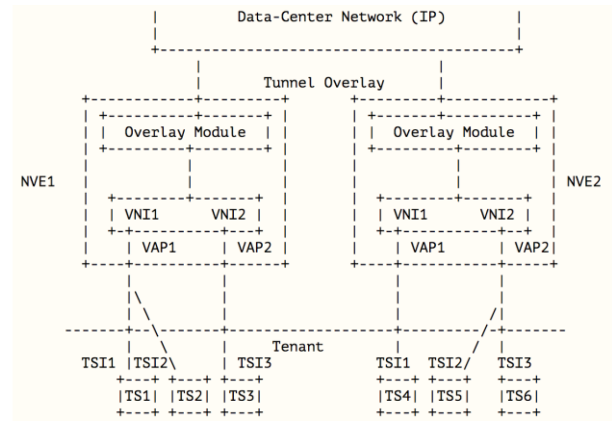


Figure 2: NVE Reference Model

RFC8014 Section 3.2 Figure 2

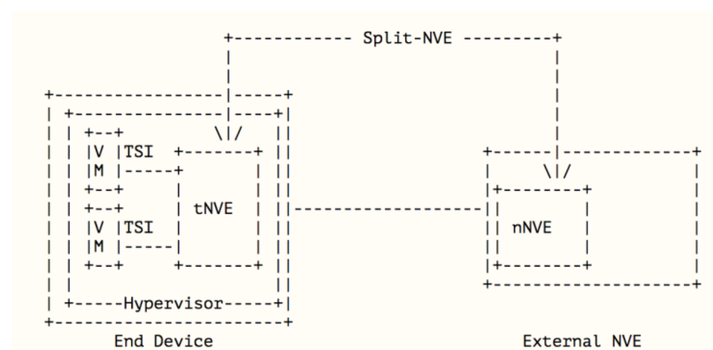


Figure 1: Split-NVE Structure

RFC8394 Section 1 Figure 1

WIP from RFC 8014

draft-bmwg-nvp-03

- RFC 8014 – Already covered in draft, just need to be consistent with naming:
  - Naming Updates:
    - Follow Terminology: NV Domain, NV Region, Tenant System Interface (TSI)
  - Additional Updates
  - Section 4 – Attach and detach state changes
    "An NVE will need to be notified when a Tenant System "attaches" to a virtual network (so it can validate the request and set up any state needed to send and receive traffic on behalf of the Tenant System on that VN). Likewise, an NVE will need to be informed when the Tenant System "detaches" from the virtual network so that it can reclaim state and resources appropriately."
  - Section 4.3 NVE State – "NVEs maintain internal data structures and state to support the sending and receiving of tenant traffic." Test Scenarios for state tracking 1-6

WIP from RFC 8394

draft-bmwg-nvp-03

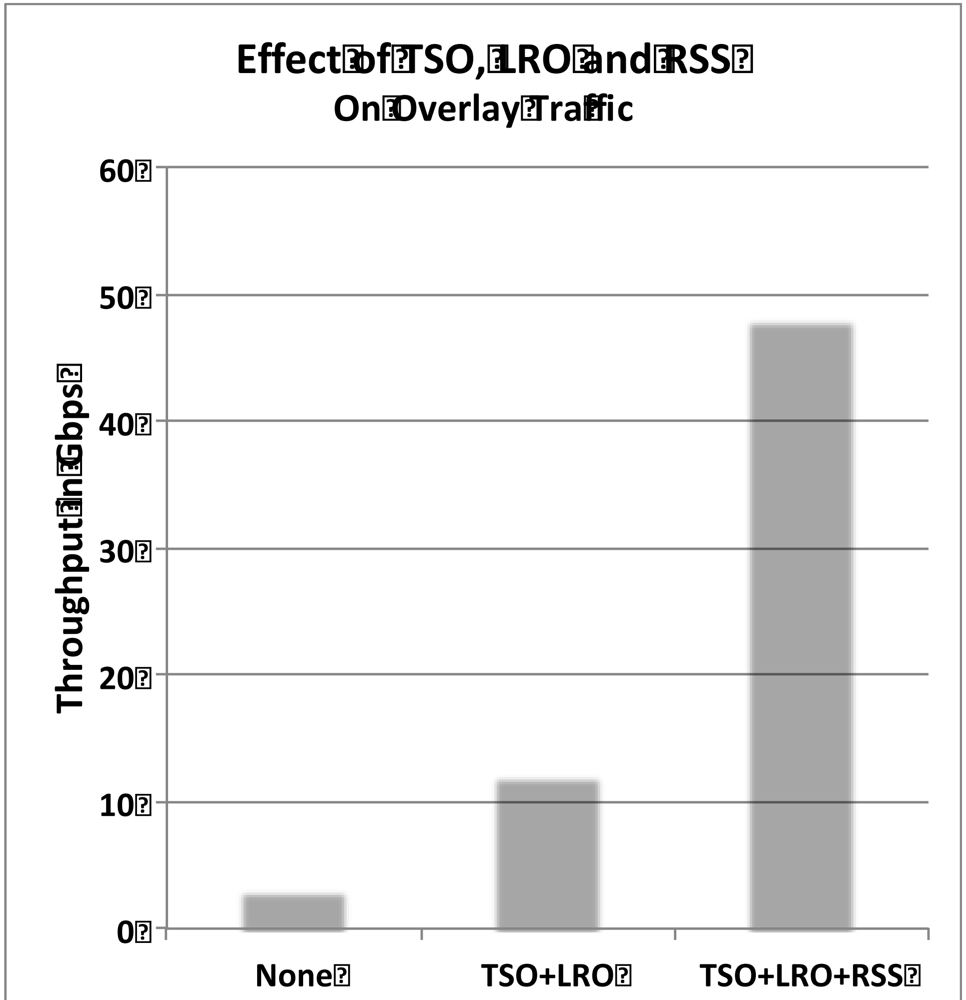
- Split NVE (RFC 8394 Section 2) – VM Lifecycle
  - Terminology update: Split-NVE, tNVE, nNVE, External NVE, VN Profile, VSI, VDP
  - State changes to VMs
    - VM Creation Event
    - VM Live Migration Event
    - VM Termination Event
    - VM Pause, Suspension, and Resumption Events
- Interactions between tNVE, nNVE and hypervisor – Example ""In the VM creation phase, the VM's TSI has to be associated with the External NVE. "Association" here indicates that the hypervisor and the External NVE have signaled each other and reached some form of agreement."

Backup Slides

Hardware Switch vs Software Switch

Hardware Switching	Logical Switch/Logical Router etc.,
Works at lower layer packets	Works closer to application layer segments
Limited by ASIC/SoC	Limited mostly by CPU and Memory (only LB) • which is not really a limit with today's processor capabilities and memory capacity/speeds
Packet size limited by supported MTU • General Max supported is 9K	Packet size a function of RSS, TSO & LRO etc., • By default 65K
Multiport – often 48 or more	Generally 2 Ports/Server
Extending functionality through additional ASIC / FPGAs and Hardware	NIC Offloads Intel DPDK / Latest Drivers etc., SSL Offload with AES-NI (Intel and AMD)

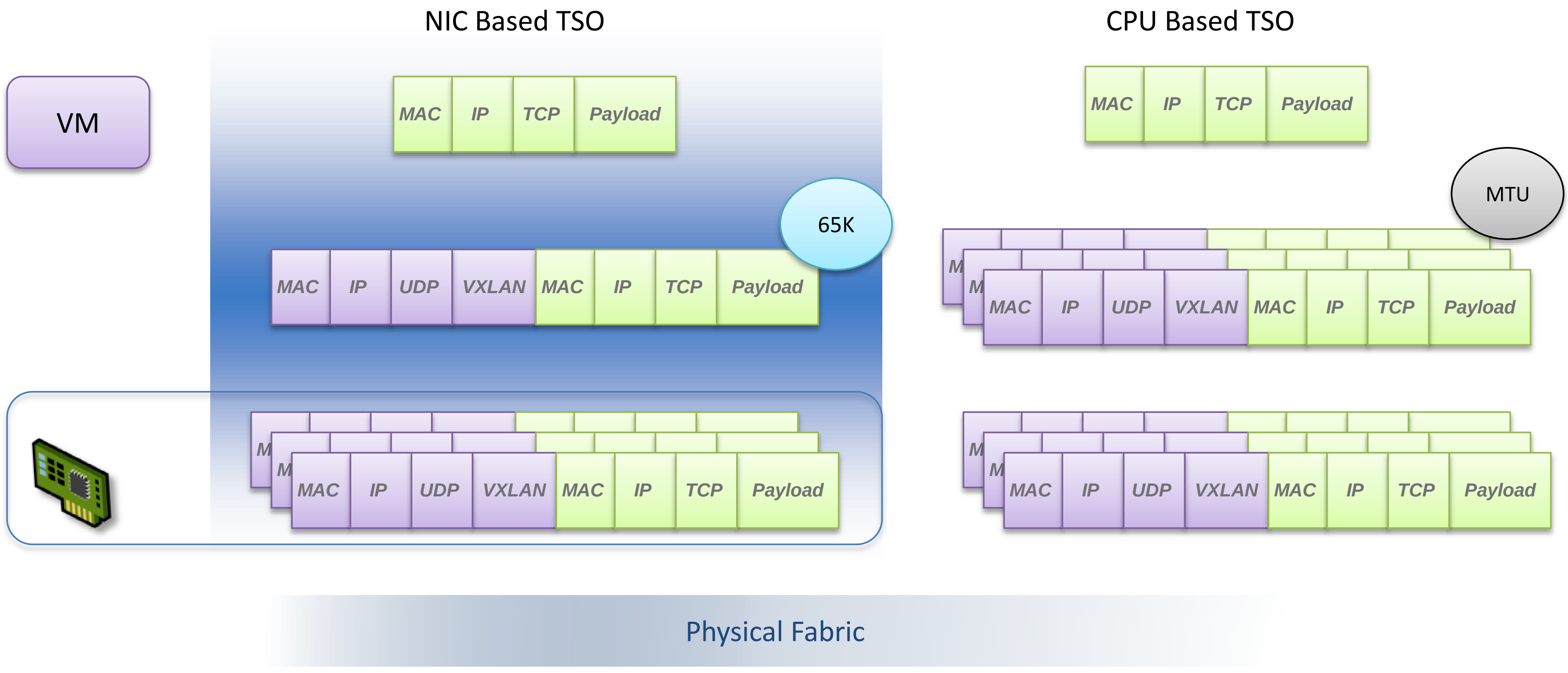
Example Results



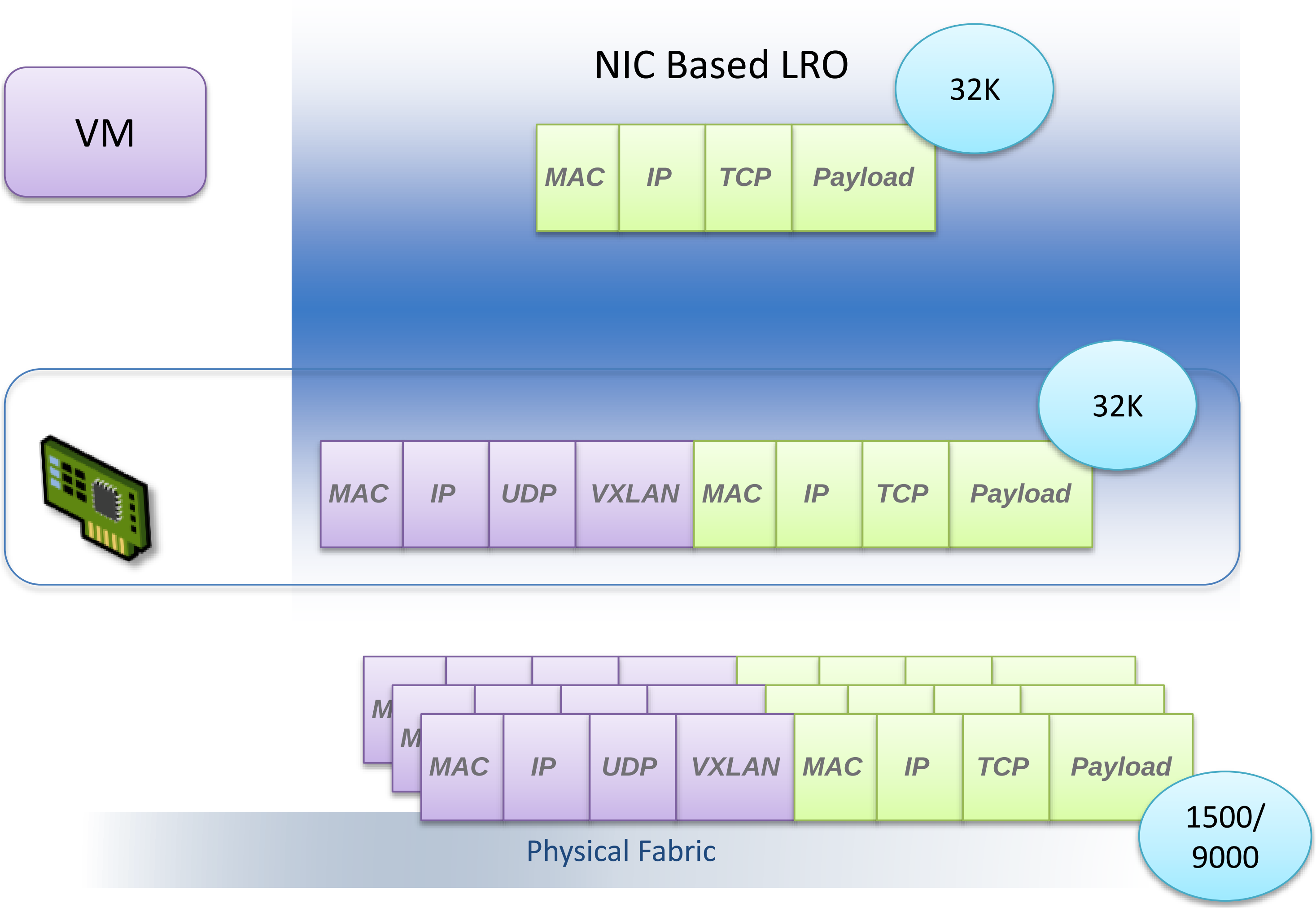
- > 10 times difference in throughput
- Throughput is a function of not just CPU but NIC card capabilities
- Other offload capabilities also have impact on performance – not profiled here
- Virtual ports don't have a rigid bandwidth profile

Server Hardware

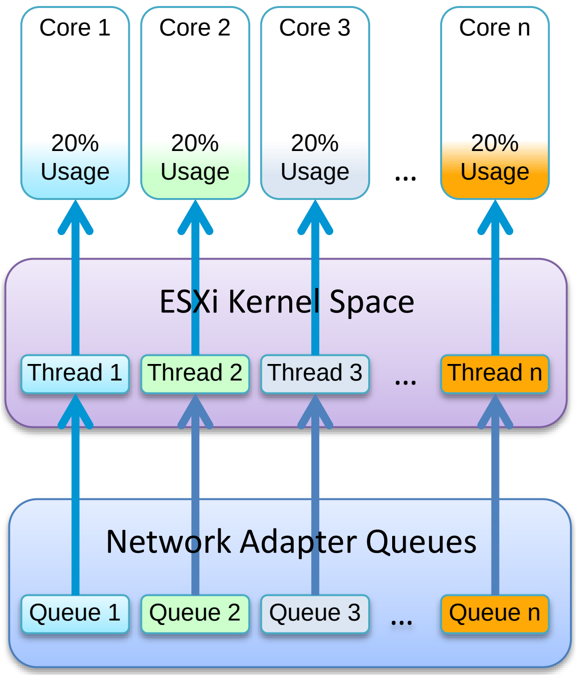
TSO for Overlay Traffic



LRO for Overlay Traffic



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

Page Size and Response Times

Average Page Size	2MB	http://httparchive.org/trends.php
Average HTML Content	56KB	
Web Response Times	200ms	https://developers.google.com/speed/docs/insights/Server
Memcached Response Time	Sub 1ms	https://code.google.com/p/memcached/wiki/NewPerformance

Documentation

Example Test Methodology

- Application level throughput using Apache Benchmark
  - ~2m file sizes based on http://httparchive.org/trends.php
    - Images tend to be larger
    - Page content tends to be smaller
- Application latency with Memslap
  - Standard settings
- iPerf
- Avalanche

Application Layer Benchmarks