Network Working Group
Internet-Draft
Intended status: Informational
Expires: April 17, 2017

M. Tahhan
B. O'Mahony
Intel
A. Morton
AT&T Labs

October 14, 2016

Benchmarking Virtual Switches in OPNFV
draft-ietf-bmwg-vswitch-opnfv-01

Abstract

This memo describes the progress of the Open Platform for NFV (OPNFV) project on virtual switch performance "VSWITCHPERF". This project intends to build on the current and completed work of the Benchmarking Methodology Working Group in IETF, by referencing existing literature. The Benchmarking Methodology Working Group has traditionally conducted laboratory characterization of dedicated physical implementations of internetworking functions. Therefore, this memo begins to describe the additional considerations when virtual switches are implemented in general-purpose hardware. The expanded tests and benchmarks are also influenced by the OPNFV mission to support virtualization of the "telco" infrastructure.

Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [RFC2119].

Status of This Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at http://datatracker.ietf.org/drafts/current/.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on April 17, 2017.

Copyright Notice

Copyright (c) 2016 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (http://trustee.ietf.org/license-info) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction 3
2. Scope 3
3. Benchmarking Considerations 4
3.1. Comparison with Physical Network Functions 4
3.2. Continued Emphasis on Black-Box Benchmarks 4
3.3. New Configuration Parameters 4
3.4. Flow classification 6
3.5. Benchmarks using Baselines with Resource Isolation . . . 7
4. VSWITCHPERF Specification Summary 8
5. 3x3 Matrix Coverage 16
5.1. Speed of Activation 17
5.2. Accuracy of Activation section 17
5.3. Reliability of Activation 17
5.4. Scale of Activation 17
5.5. Speed of Operation 17
5.6. Accuracy of Operation 17
5.7. Reliability of Operation 17
5.8. Scalability of Operation 18
5.9. Summary 18
6. Security Considerations 18
7. IANA Considerations 19
8. Acknowledgements 19
9. References 19
9.1. Normative References 19
9.2. Informative References 21
Authors' Addresses 22

1. Introduction

Benchmarking Methodology Working Group (BMWG) has traditionally conducted laboratory characterization of dedicated physical implementations of internetworking functions. The Black-box Benchmarks of Throughput, Latency, Forwarding Rates and others have served our industry for many years. Now, Network Function Virtualization (NFV) has the goal to transform how internetwork functions are implemented, and therefore has garnered much attention.

This memo summarizes the progress of the Open Platform for NFV (OPNFV) project on virtual switch performance characterization, "VSWITCHPERF", through the Brahmaputra (second) release [BrahRel]. This project intends to build on the current and completed work of the Benchmarking Methodology Working Group in IETF, by referencing existing literature. For example, currently the most often referenced RFC is [RFC2544] (which depends on [RFC1242]) and foundation of the benchmarking work in OPNFV is common and strong.

See https://wiki.opnfv.org/characterize_vswitch_performance_for_telco_nfv_use_cases for more background, and the OPNFV website for general information: https://www.opnfv.org/

The authors note that OPNFV distinguishes itself from other open source compute and networking projects through its emphasis on existing "telco" services as opposed to cloud-computing. There are many ways in which telco requirements have different emphasis on performance dimensions when compared to cloud computing: support for and transfer of isochronous media streams is one example.

Note also that the move to NFV Infrastructure has resulted in many new benchmarking initiatives across the industry. The authors are currently doing their best to maintain alignment with many other projects, and this Internet Draft is one part of the efforts. We acknowledge the early work in [I-D.huang-bmwg-virtual-network-performance], and useful discussion with the authors.

2. Scope

The primary purpose and scope of the memo is to inform the industry of work-in-progress that builds on the body of extensive BMWG literature and experience, and describe the extensions needed for benchmarking virtual switches. Inital feedback indicates that many of these extensions may be applicable beyond the current scope (to hardware switches in the NFV Infrastructure and to virtual routers, for example). Additionally, this memo serves as a vehicle to include

more detail and commentary from BMWG and other Open Source communities, under BMWG's chartered work to characterize the NFV Infrastructure (a virtual switch is an important aspect of that infrastructure).

The benchmarking covered in this memo should be applicable to many types of vswitches, and remain vswitch-agnostic to great degree. There has been no attempt to track and test all features of any specific vswitch implementation.

3. Benchmarking Considerations

This section highlights some specific considerations (from [I-D.ietf-bmwg-virtual-net])related to Benchmarks for virtual switches. The OPNFV project is sharing its present view on these areas, as they develop their specifications in the Level Test Design (LTD) document.

3.1. Comparison with Physical Network Functions

To compare the performance of virtual designs and implementations with their physical counterparts, identical benchmarks are needed. BMWG has developed specifications for many network functions this memo re-uses existing benchmarks through references, and expands them during development of new methods. A key configuration aspect is the number of parallel cores required to achieve comparable performance with a given physical device, or whether some limit of scale was reached before the cores could achieve the comparable level.

It's unlikely that the virtual switch will be the only application running on the SUT, so CPU utilization, Cache utilization, and Memory footprint should also be recorded for the virtual implementations of internetworking functions.

3.2. Continued Emphasis on Black-Box Benchmarks

External observations remain essential as the basis for Benchmarks. Internal observations with fixed specification and interpretation will be provided in parallel to assist the development of operations procedures when the technology is deployed.

3.3. New Configuration Parameters

A key consideration when conducting any sort of benchmark is trying to ensure the consistency and repeatability of test results. When benchmarking the performance of a vSwitch there are many factors that can affect the consistency of results, one key factor is matching the various hardware and software details of the SUT. This section lists

some of the many new parameters which this project believes are critical to report in order to achieve repeatability.

Hardware details including:

- Platform details
- Processor details
- Memory information (type and size)
- Number of enabled cores
- Number of cores used for the test
- Number of physical NICs, as well as their details (manufacturer, versions, type and the PCI slot they are plugged into)
- NIC interrupt configuration
- BIOS version, release date and any configurations that were modified
- CPU microcode level
- Memory DIMM configurations (quad rank performance may not be the same as dual rank) in size, freq and slot locations
- PCI configuration parameters (payload size, early ack option...)
- Power management at all levels (ACPI sleep states, processor package, OS...)

Software details including:

- OS parameters and behavior (text vs graphical no one typing at the console on one system)
- OS version (for host and VNF)
- Kernel version (for host and VNF)
- GRUB boot parameters (for host and VNF)
- Hypervisor details (Type and version)
- Selected vSwitch, version number or commit id used

- vSwitch launch command line if it has been parameterised
- Memory allocation to the vSwitch
- which NUMA node it is using, and how many memory channels
- DPDK or any other SW dependency version number or commit id used
- Memory allocation to a VM - if it's from Hugpages/elsewhere
- VM storage type: snapshot/independent persistent/independent non-persistent
- Number of VMs
- Number of Virtual NICs (vNICs), versions, type and driver
- Number of virtual CPUs and their core affinity on the host
- Number vNIC interrupt configuration
- Thread affinitization for the applications (including the vSwitch itself) on the host
- Details of Resource isolation, such as CPUs designated for Host/ Kernel (isolcpu) and CPUs designated for specific processes (taskset). - Test duration. - Number of flows.

Test Traffic Information:

- Traffic type - UDP, TCP, IMIX / Other
- Packet Sizes
- Deployment Scenario

3.4. Flow classification

Virtual switches group packets into flows by processing and matching particular packet or frame header information, or by matching packets based on the input ports. Thus a flow can be thought of a sequence of packets that have the same set of header field values (5-tuple) or have arrived on the same port. Performance results can vary based on the parameters the vSwitch uses to match for a flow. The recommended flow classification parameters for any vSwitch performance tests are: the input port, the source IP address, the destination IP address and the Ethernet protocol type field. It is essential to increase the flow timeout time on a vSwitch before conducting any performance

tests that do not measure the flow setup time. Normally the first packet of a particular stream will install the flow in the virtual switch which adds an additional latency, subsequent packets of the same flow are not subject to this latency if the flow is already installed on the vSwitch.

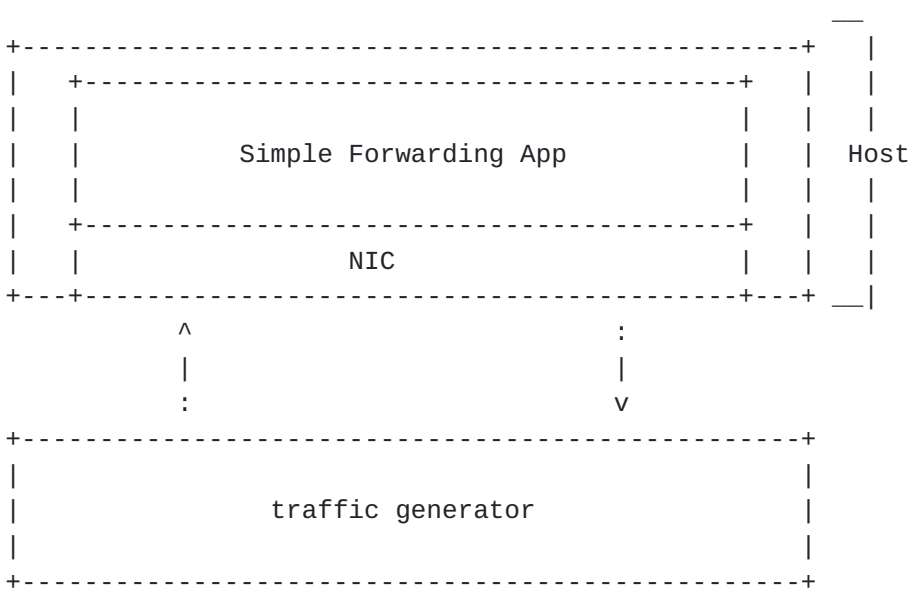
3.5. Benchmarks using Baselines with Resource Isolation

This outline describes measurement of baseline with isolated resources at a high level, which is the intended approach at this time.

1. Baselines:

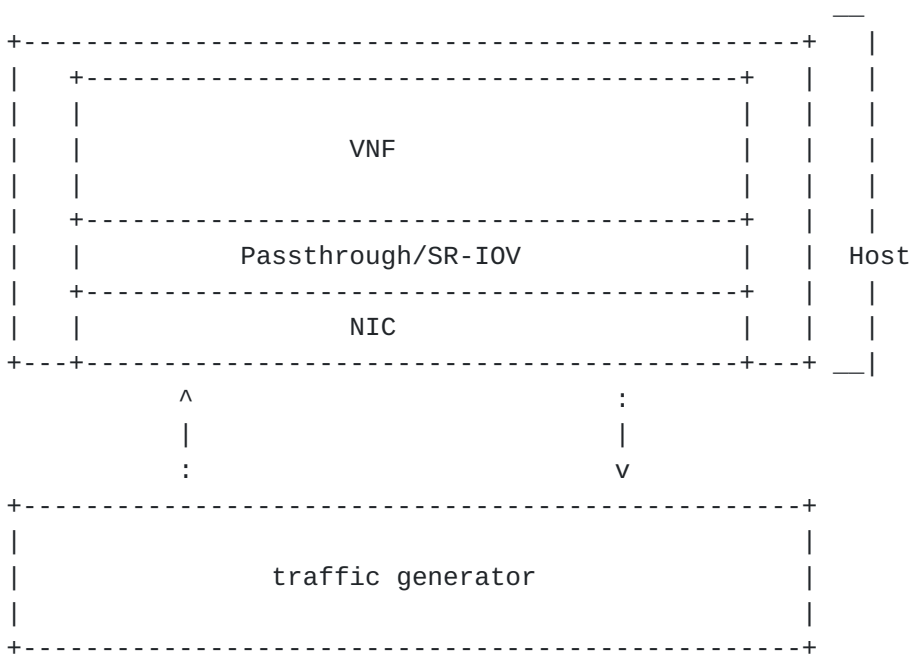
   * Optional: Benchmark platform forwarding capability without a vswitch or VNF for at least 72 hours (serves as a means of platform validation and a means to obtain the base performance for the platform in terms of its maximum forwarding rate and latency).

Benchmark platform forwarding capability



   * Benchmark VNF forwarding capability with direct connectivity (vSwitch bypass, e.g., SR/IOV) for at least 72 hours (serves as a means of VNF validation and a means to obtain the base performance for the VNF in terms of its maximum forwarding rate and latency). The metrics gathered from this test will serve as a key comparison point for vSwitch bypass technologies performance and vSwitch performance.

Benchmark VNF forwarding capability



* Benchmarking with isolated resources alone, with other resources (both HW&SW) disabled Example, vSw and VM are SUT

* Benchmarking with isolated resources alone, leaving some resources unused

* Benchmark with isolated resources and all resources occupied

2. Next Steps

* Limited sharing

* Production scenarios

* Stressful scenarios

4. VSWITCHPERF Specification Summary

The overall specification in preparation is referred to as a Level Test Design (LTD) document, which will contain a suite of performance tests. The base performance tests in the LTD are based on the pre-existing specifications developed by BMWG to test the performance of physical switches. These specifications include:

o [RFC2544] Benchmarking Methodology for Network Interconnect Devices

- [RFC2889] Benchmarking Methodology for LAN Switching
- [RFC6201] Device Reset Characterization
- [RFC5481] Packet Delay Variation Applicability Statement

Some of the above/newer RFCs are being applied in benchmarking for the first time, and represent a development challenge for test equipment developers. Fortunately, many members of the testing system community have engaged on the VSPERF project, including an open source test system.

In addition to this, the LTD also re-uses the terminology defined by:

- [RFC2285] Benchmarking Terminology for LAN Switching Devices
- [RFC5481] Packet Delay Variation Applicability Statement

Specifications to be included in future updates of the LTD include:

- [RFC3918] Methodology for IP Multicast Benchmarking
- [RFC4737] Packet Reordering Metrics

As one might expect, the most fundamental internetworking characteristics of Throughput and Latency remain important when the switch is virtualized, and these benchmarks figure prominently in the specification.

When considering characteristics important to "telco" network functions, we must begin to consider additional performance metrics. In this case, the project specifications have referenced metrics from the IETF IP Performance Metrics (IPPM) literature. This means that the [RFC2544] test of Latency is replaced by measurement of a metric derived from IPPM's [RFC2679], where a set of statistical summaries will be provided (mean, max, min, etc.). Further metrics planned to be benchmarked include packet delay variation as defined by [RFC5481] , reordering, burst behaviour, DUT availability, DUT capacity and packet loss in long term testing at Throughput level, where some low-level of background loss may be present and characterized.

Tests have been (or will be) designed to collect the metrics below:

- Throughput Tests to measure the maximum forwarding rate (in frames per second or fps) and bit rate (in Mbps) for a constant load (as defined by [RFC1242]) without traffic loss.

- Packet and Frame Delay Distribution Tests to measure average, min and max packet and frame delay for constant loads.

- Packet Delay Tests to understand latency distribution for different packet sizes and over an extended test run to uncover outliers.

- Scalability Tests to understand how the virtual switch performs as the number of flows, active ports, complexity of the forwarding logic's configuration... it has to deal with increases.

- Stream Performance Tests (TCP, UDP) to measure bulk data transfer performance, i.e. how fast systems can send and receive data through the switch.

- Control Path and Datapath Coupling Tests, to understand how closely coupled the datapath and the control path are as well as the effect of this coupling on the performance of the DUT (example: delay of the initial packet of a flow).

- CPU and Memory Consumption Tests to understand the virtual switch's footprint on the system, usually conducted as auxiliary measurements with benchmarks above. They include: CPU utilization, Cache utilization and Memory footprint.

- The so-called "Soak" tests, where the selected test is conducted over a long period of time (with an ideal duration of 24 hours, but only long enough to determine that stability issues exist when found; there is no requirement to continue a test when a DUT exhibits instability over time). The key performance characteristics and benchmarks for a DUT are determined (using short duration tests) prior to conducting soak tests. The purpose of soak tests is to capture transient changes in performance which may occur due to infrequent processes, memory leaks, or the low probability coincidence of two or more processes. The stability of the DUT is the paramount consideration, so performance must be evaluated periodically during continuous testing, and this results in use of [RFC2889] Frame Rate metrics instead of [RFC2544] Throughput (which requires stopping traffic to allow time for all traffic to exit internal queues), for example.

Future/planned test specs include:

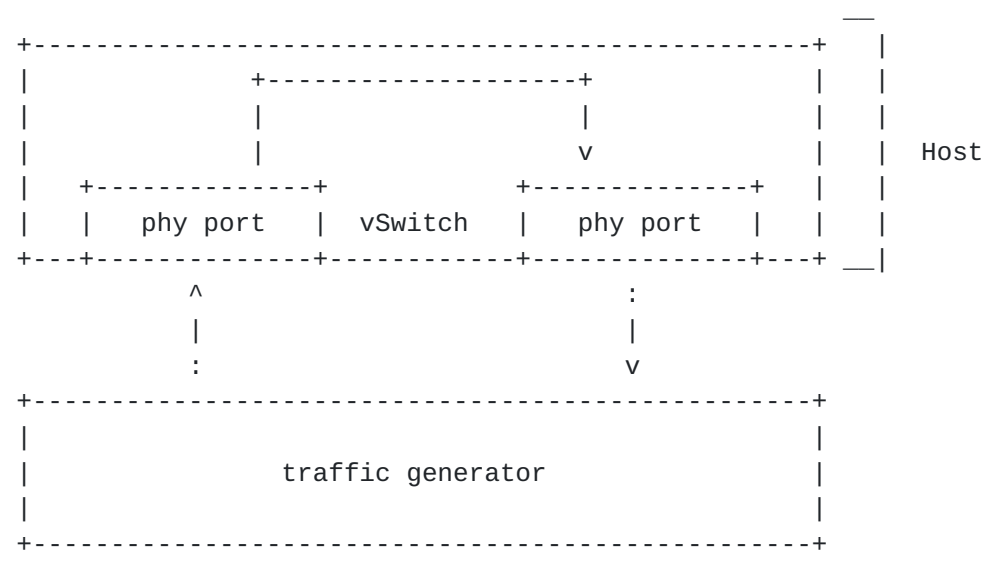
- Request/Response Performance Tests (TCP, UDP) which measure the transaction rate through the switch.

- Noisy Neighbour Tests, to understand the effects of resource sharing on the performance of a virtual switch.

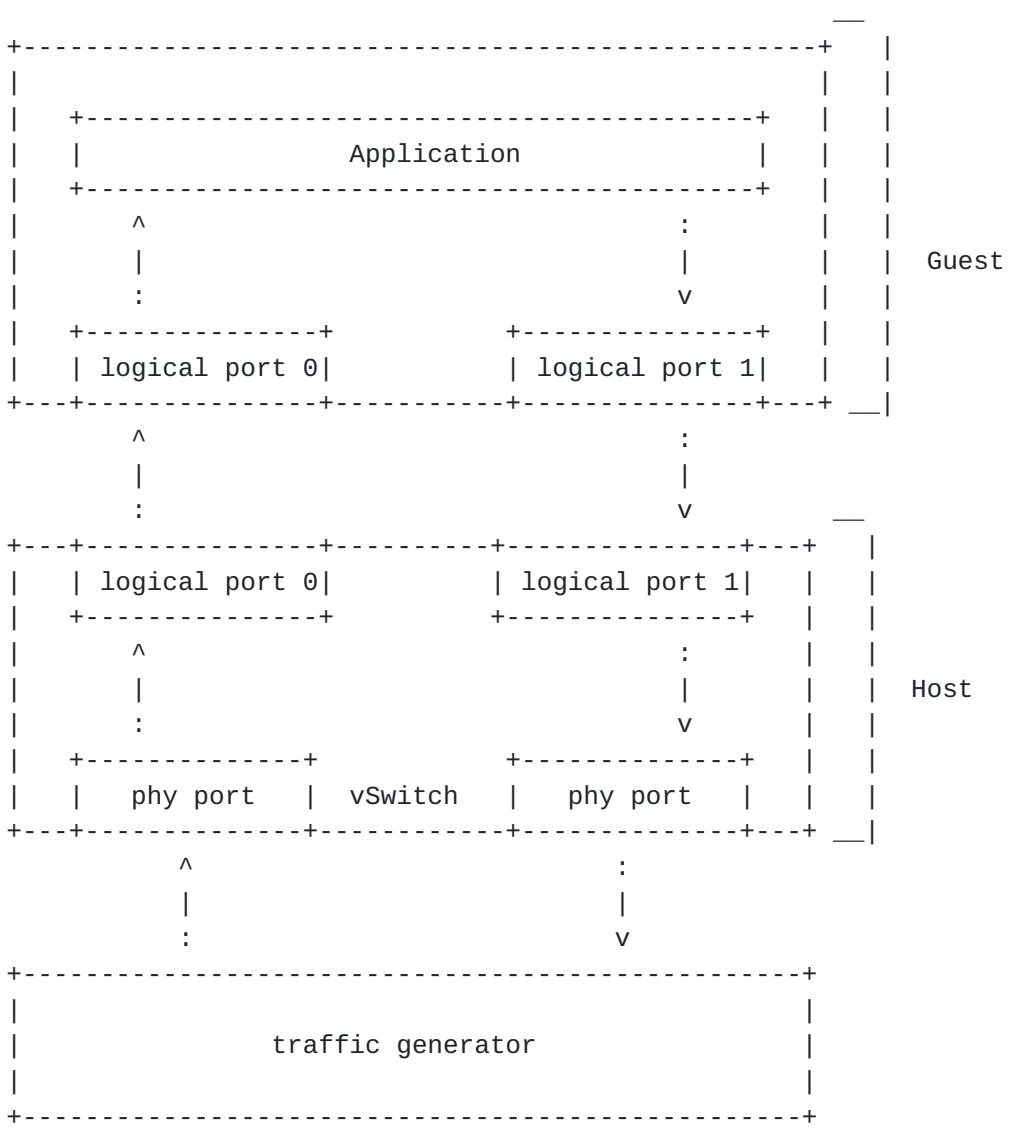
o Tests derived from examination of ETSI NFV Draft GS IFA003 requirements [IFA003] on characterization of acceleration technologies applied to vswitches.

The flexibility of deployment of a virtual switch within a network means that the BMWG IETF existing literature needs to be used to characterize the performance of a switch in various deployment scenarios. The deployment scenarios under consideration include:

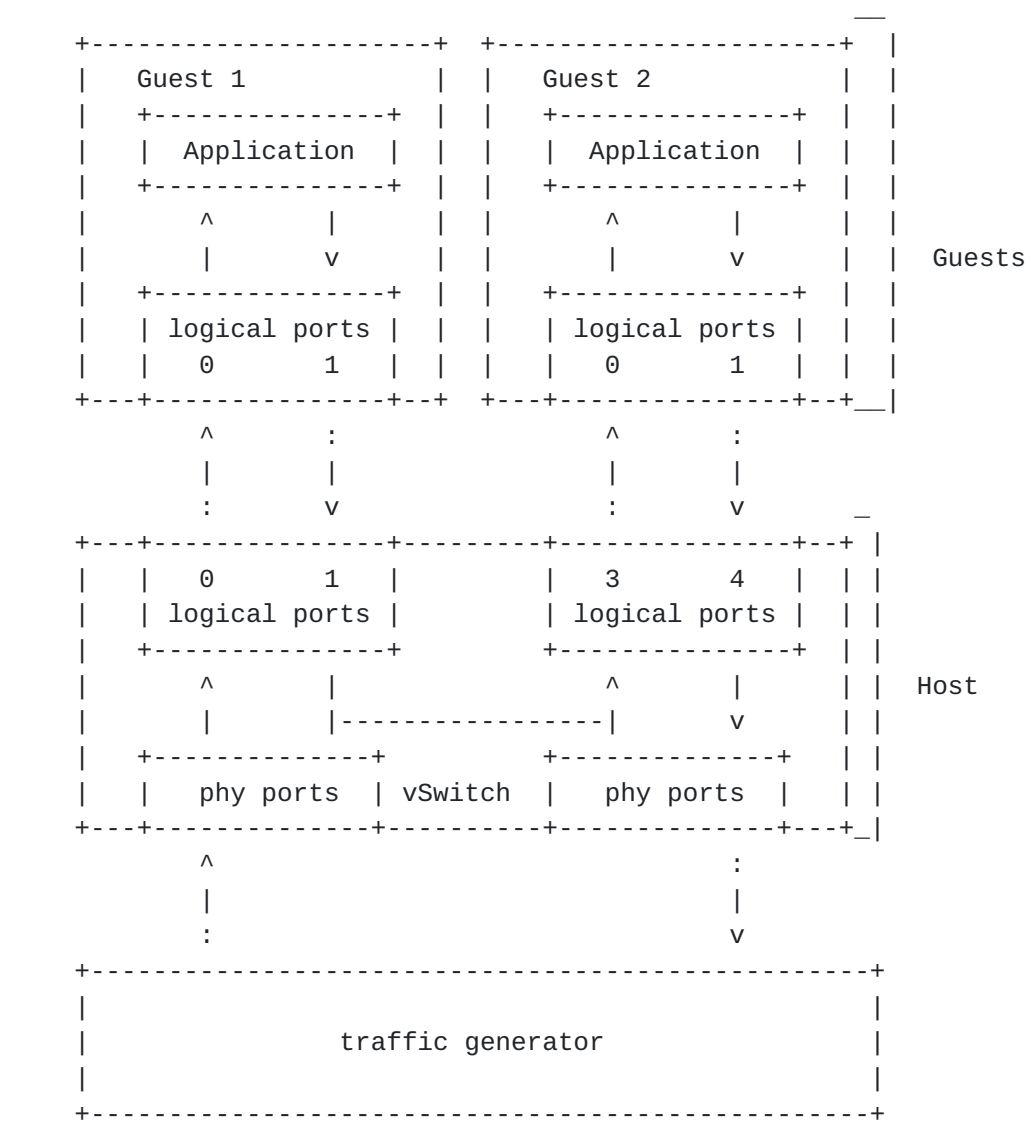
Physical port to virtual switch to physical port



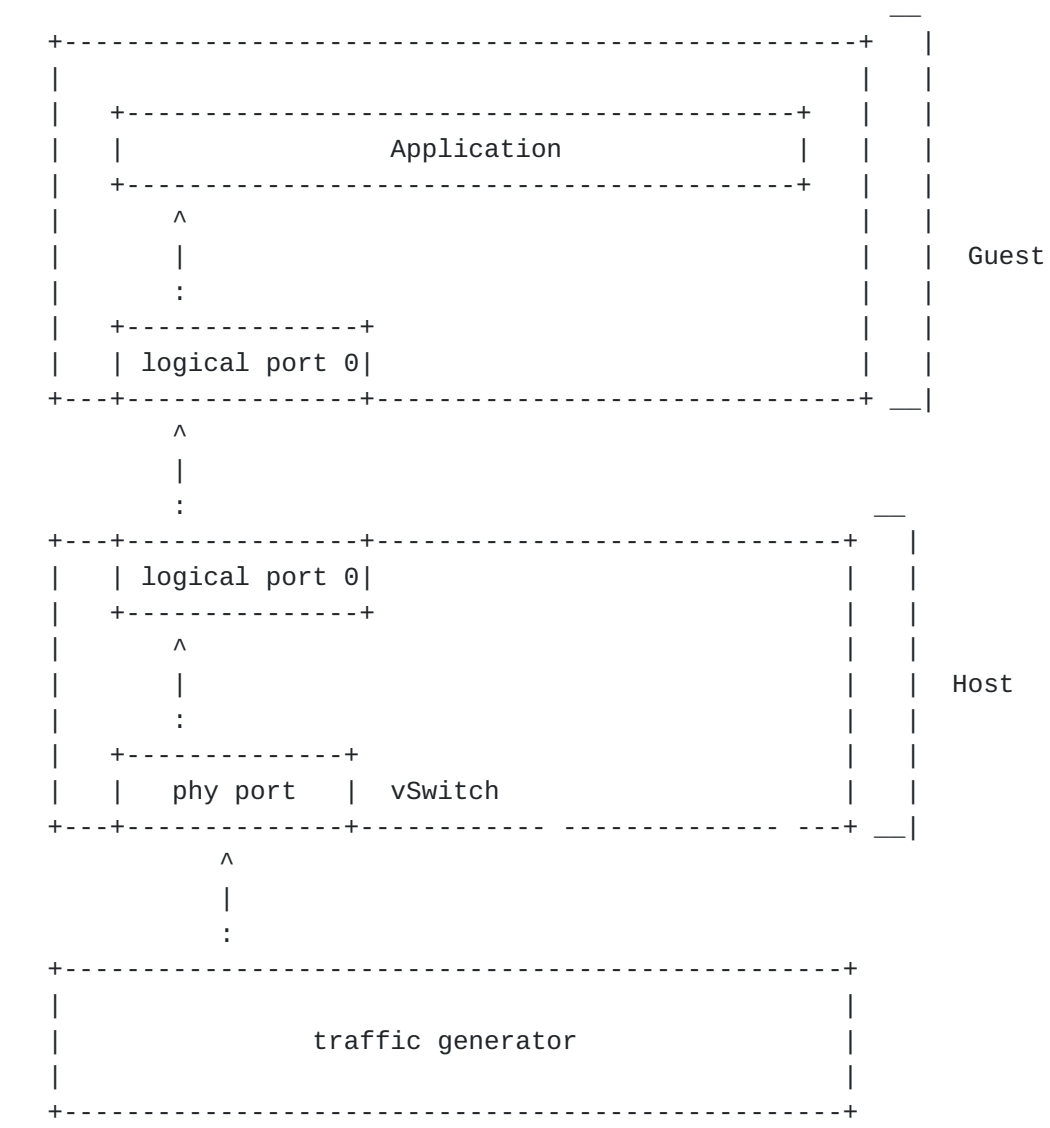
Physical port to virtual switch to VNF to virtual switch to physical port



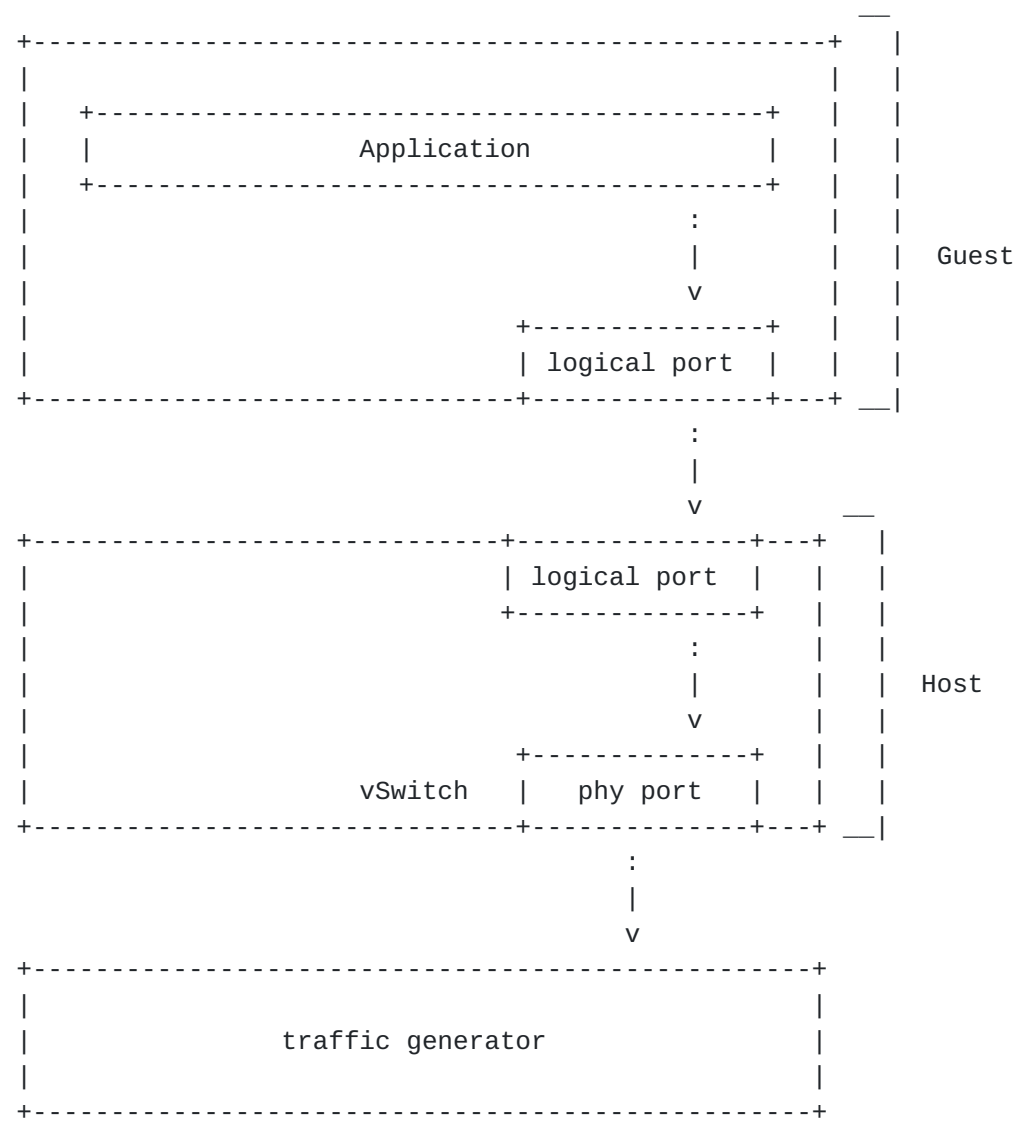
Physical port to virtual switch to VNF to virtual switch to VNF to virtual switch to physical port



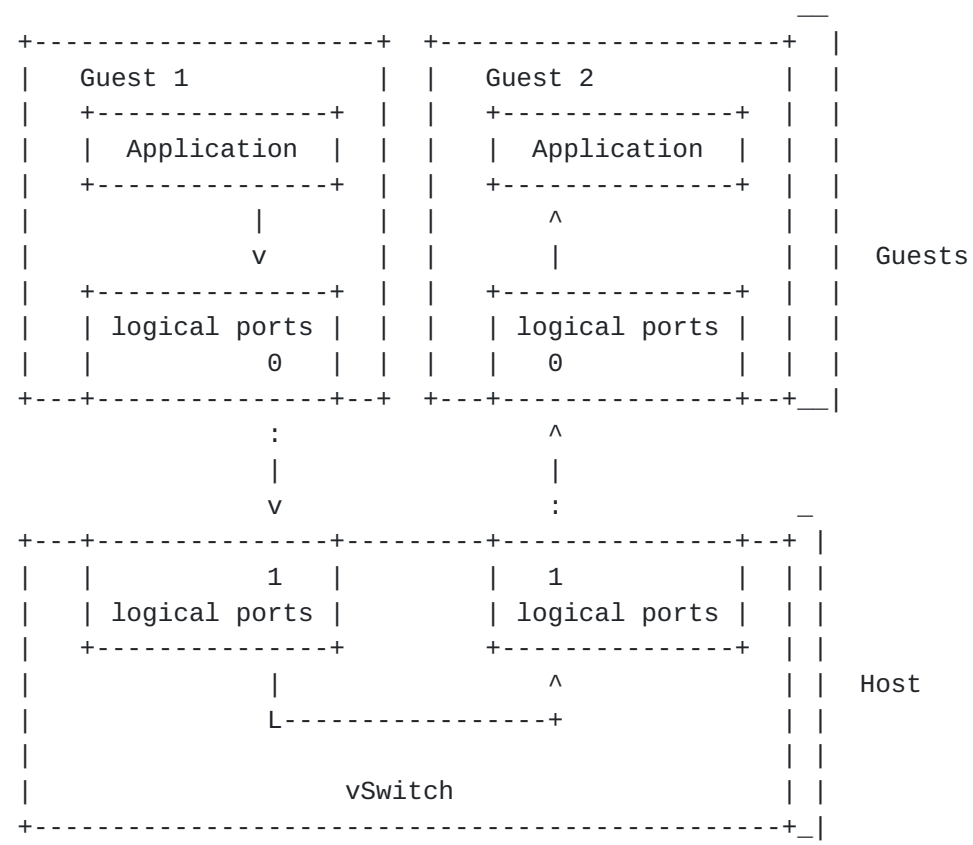
Physical port to virtual switch to VNF



VNF to virtual switch to physical port



VNF to virtual switch to VNF



A set of Deployment Scenario figures is available on the VSPERF Test Methodology Wiki page [TestTopo].

5. 3x3 Matrix Coverage

This section organizes the many existing test specifications into the "3x3" matrix (introduced in [I-D.ietf-bmwg-virtual-net]). Because the LTD specification ID names are quite long, this section is organized into lists for each occupied cell of the matrix (not all are occupied, also the matrix has grown to 3x4 to accommodate scale metrics when displaying the coverage of many metrics/benchmarks). The current version of the LTD specification is available [LTD].

The tests listed below assess the activation of paths in the data plane, rather than the control plane.

A complete list of tests with short summaries is available on the VSPERF "LTD Test Spec Overview" Wiki page [LTDoverV].

5.1. Speed of Activation

- Activation.RFC2889.AddressLearningRate
- PacketLatency.InitialPacketProcessingLatency

5.2. Accuracy of Activation section

- CPDP.Coupling.Flow.Addition

5.3. Reliability of Activation

- Throughput.RFC2544.SystemRecoveryTime
- Throughput.RFC2544.ResetTime

5.4. Scale of Activation

- Activation.RFC2889.AddressCachingCapacity

5.5. Speed of Operation

- Throughput.RFC2544.PacketLossRate
- CPU.RFC2544.0PacketLoss
- Throughput.RFC2544.PacketLossRateFrameModification
- Throughput.RFC2544.BackToBackFrames
- Throughput.RFC2889.MaxForwardingRate
- Throughput.RFC2889.ForwardPressure
- Throughput.RFC2889.BroadcastFrameForwarding

5.6. Accuracy of Operation

- Throughput.RFC2889.ErrorFramesFiltering
- Throughput.RFC2544.Profile

5.7. Reliability of Operation

- Throughput.RFC2889.Soak
- Throughput.RFC2889.SoakFrameModification

- PacketDelayVariation.RFC3393.Soak

5.8. Scalability of Operation

- Scalability.RFC2544.0PacketLoss

- MemoryBandwidth.RFC2544.0PacketLoss.Scalability

5.9. Summary

	SPEED	ACCURACY	RELIABILITY	SCALE
Activation	X	X	X	X
Operation	X	X	X	X
De-activation				

6. Security Considerations

Benchmarking activities as described in this memo are limited to technology characterization of a Device Under Test/System Under Test (DUT/SUT) using controlled stimuli in a laboratory environment, with dedicated address space and the constraints specified in the sections above.

The benchmarking network topology will be an independent test setup and MUST NOT be connected to devices that may forward the test traffic into a production network, or misroute traffic to the test management network.

Further, benchmarking is performed on a "black-box" basis, relying solely on measurements observable external to the DUT/SUT.

Special capabilities SHOULD NOT exist in the DUT/SUT specifically for benchmarking purposes. Any implications for network security arising from the DUT/SUT SHOULD be identical in the lab and in production networks.

7. IANA Considerations

No IANA Action is requested at this time.

8. Acknowledgements

The authors appreciate and acknowledge comments from Scott Bradner, Marius Georgescu, Ramki Krishnan, Doug Montgomery, Martin Klozik, Christian Trautman, and others for their reviews.

9. References

9.1. Normative References

[NFV.PER001] "Network Function Virtualization: Performance and Portability Best Practices", Group Specification ETSI GS NFV-PER 001 V1.1.1 (2014-06), June 2014.

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <http://www.rfc-editor.org/info/rfc2119>.

[RFC2285] Mandeville, R., "Benchmarking Terminology for LAN Switching Devices", RFC 2285, DOI 10.17487/RFC2285, February 1998, <http://www.rfc-editor.org/info/rfc2285>.

[RFC2330] Paxson, V., Almes, G., Mahdavi, J., and M. Mathis, "Framework for IP Performance Metrics", RFC 2330, DOI 10.17487/RFC2330, May 1998, <http://www.rfc-editor.org/info/rfc2330>.

[RFC2544] Bradner, S. and J. McQuaid, "Benchmarking Methodology for Network Interconnect Devices", RFC 2544, DOI 10.17487/RFC2544, March 1999, <http://www.rfc-editor.org/info/rfc2544>.

[RFC2679] Almes, G., Kalidindi, S., and M. Zekauskas, "A One-way Delay Metric for IPPM", RFC 2679, DOI 10.17487/RFC2679, September 1999, <http://www.rfc-editor.org/info/rfc2679>.

[RFC2680] Almes, G., Kalidindi, S., and M. Zekauskas, "A One-way Packet Loss Metric for IPPM", RFC 2680, DOI 10.17487/RFC2680, September 1999, <http://www.rfc-editor.org/info/rfc2680>.

[RFC2681] Almes, G., Kalidindi, S., and M. Zekauskas, "A Round-trip Delay Metric for IPPM", RFC 2681, DOI 10.17487/RFC2681, September 1999, <http://www.rfc-editor.org/info/rfc2681>.

[RFC2889] Mandeville, R. and J. Perser, "Benchmarking Methodology for LAN Switching Devices", RFC 2889, DOI 10.17487/RFC2889, August 2000, <http://www.rfc-editor.org/info/rfc2889>.

[RFC3393] Demichelis, C. and P. Chimento, "IP Packet Delay Variation Metric for IP Performance Metrics (IPPM)", RFC 3393, DOI 10.17487/RFC3393, November 2002, <http://www.rfc-editor.org/info/rfc3393>.

[RFC3432] Raisanen, V., Grotefeld, G., and A. Morton, "Network performance measurement with periodic streams", RFC 3432, DOI 10.17487/RFC3432, November 2002, <http://www.rfc-editor.org/info/rfc3432>.

[RFC3918] Stopp, D. and B. Hickman, "Methodology for IP Multicast Benchmarking", RFC 3918, DOI 10.17487/RFC3918, October 2004, <http://www.rfc-editor.org/info/rfc3918>.

[RFC4689] Poretsky, S., Perser, J., Erramilli, S., and S. Khurana, "Terminology for Benchmarking Network-layer Traffic Control Mechanisms", RFC 4689, DOI 10.17487/RFC4689, October 2006, <http://www.rfc-editor.org/info/rfc4689>.

[RFC4737] Morton, A., Ciavattone, L., Ramachandran, G., Shalunov, S., and J. Perser, "Packet Reordering Metrics", RFC 4737, DOI 10.17487/RFC4737, November 2006, <http://www.rfc-editor.org/info/rfc4737>.

[RFC5357] Hedayat, K., Krzanowski, R., Morton, A., Yum, K., and J. Babiarz, "A Two-Way Active Measurement Protocol (TWAMP)", RFC 5357, DOI 10.17487/RFC5357, October 2008, <http://www.rfc-editor.org/info/rfc5357>.

[RFC5905] Mills, D., Martin, J., Ed., Burbank, J., and W. Kasch, "Network Time Protocol Version 4: Protocol and Algorithms Specification", RFC 5905, DOI 10.17487/RFC5905, June 2010, <http://www.rfc-editor.org/info/rfc5905>.

[RFC6201] Asati, R., Pignataro, C., Calabria, F., and C. Olvera, "Device Reset Characterization", RFC 6201, DOI 10.17487/RFC6201, March 2011, <http://www.rfc-editor.org/info/rfc6201>.

9.2. Informative References

[BrahRel] "Brahmaputra, Second OPNFV Release https://www.opnfv.org/brahmaputra".

[I-D.huang-bmwg-virtual-network-performance] Huang, L., Rong, G., Mandeville, B., and B. Hickman, "Benchmarking Methodology for Virtualization Network Performance", draft-huang-bmwg-virtual-network-performance-01 (work in progress), April 2015.

[I-D.ietf-bmwg-virtual-net] Morton, A., "Considerations for Benchmarking Virtual Network Functions and Their Infrastructure", draft-ietf-bmwg-virtual-net-04 (work in progress), August 2016.

[IFA003] "https://docbox.etsi.org/ISG/NFV/Open/Drafts/IFA003_Acceleration_-_vSwitch_Spec/".

[LTD] "LTD Test Specification http://artifacts.opnfv.org/vswitchperf/brahmaputra/docs/requirements/index.html".

[LTDoverV] "LTD Test Spec Overview https://wiki.opnfv.org/wiki/vswitchperf_test_spec_review".

[RFC1242] Bradner, S., "Benchmarking Terminology for Network Interconnection Devices", RFC 1242, DOI 10.17487/RFC1242, July 1991, <http://www.rfc-editor.org/info/rfc1242>.

[RFC5481] Morton, A. and B. Claise, "Packet Delay Variation Applicability Statement", RFC 5481, DOI 10.17487/RFC5481, March 2009, <http://www.rfc-editor.org/info/rfc5481>.

[RFC6049] Morton, A. and E. Stephan, "Spatial Composition of Metrics", RFC 6049, DOI 10.17487/RFC6049, January 2011, <http://www.rfc-editor.org/info/rfc6049>.

[RFC6248] Morton, A., "RFC 4148 and the IP Performance Metrics (IPPM) Registry of Metrics Are Obsolete", RFC 6248, DOI 10.17487/RFC6248, April 2011, <http://www.rfc-editor.org/info/rfc6248>.

[RFC6390] Clark, A. and B. Claise, "Guidelines for Considering New Performance Metric Development", BCP 170, RFC 6390, DOI 10.17487/RFC6390, October 2011, <http://www.rfc-editor.org/info/rfc6390>.

[TestTopo]
"Test Topologies https://wiki.opnfv.org/vsperf/test_methodology".

Authors' Addresses

Maryam Tahhan
Intel

Email: maryam.tahhan@intel.com

Billy O'Mahony
Intel

Email: billy.o.mahony@intel.com

Al Morton
AT&T Labs
200 Laurel Avenue South
Middletown,, NJ 07748
USA

Phone: +1 732 420 1571
Fax: +1 732 368 1192
Email: acmorton@att.com
URI: http://home.comcast.net/~acmacm/