Internet Engineering Task Force
Internet-Draft
Intended status: Informational
Expires: May 10,2014

L. Avramov
Cisco Systems
J. Rapp
Hewlett Packard

November 6, 2013

Data Center Benchmarking Methodology
draft-bmwg-dcbench-methodology-02

Abstract

The purpose of this informational document is to establish test and evaluation methodology and measurement techniques for network equipment in the data center.

Status of this Memo

Copyright Notice

Table of Contents

1. Introduction

Traffic patterns in the data center are not uniform and are constantly changing. They are dictated by the nature and variety of applications utilized in the data center. It can be largely east-west traffic flows in one data center and north-south in another, while some may combine both. Traffic patterns can be bursty in nature and contain many-to-one, many-to-many, or one-to-many flows. Each flow may also be small and latency sensitive or large and throughput sensitive while containing a mix of UDP and TCP traffic. All of which can coexist in a single cluster and flow through a single network device all at the same time. Benchmarking of network devices have long used RFC1242, RFC2432, RFC2544, RFC2889 and RFC3918. These benchmarks have largely been focused around various latency attributes and max throughput of the Device Under Test [DUT] being benchmarked. These standards are good at measuring theoretical max throughput, forwarding rates and latency under testing conditions however, they do not represent real traffic patterns that may affect these networking devices.

The following provides a methodology for benchmarking Data Center DUT including congestion scenarios, switch buffer analysis, microburst, head of line blocking, while also using a wide mix of traffic conditions.

1.1. 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 [6].

1.2. Methodology format

The format used for each section of this document is the following:

-Objective

-Methodology

-Reporting Format

MUST: minimum test for the scenario described

SHOULD: recommended test for the scenario described

MAY: ideal test for the scenario described

2. Line Rate Testing

2.1 Objective

Provide at maximum rate test for the performance values for throughput, latency and jitter. It is meant to provide the tests to run and methodology to verify that a DUT is capable of forwarding packets at line rate under non-congested conditions.

2.2 Methodology

A traffic generator SHOULD be connected to all ports on the DUT. Two tests MUST be conducted: a port-pair test [RFC 2544/3918 compliant] and also in a full mesh type of DUT test [RFC 2889/3918 compliant].

For all tests, the percentage of traffic per port capacity sent MUST be 99.98% at most, with no PPM adjustment to ensure stressing the DUT in worst case conditions. Tests results at a lower rate MAY be provided for better understanding of performance increase in terms of latency and jitter when the rate is lower than 99.98%. The receiving rate of the traffic needs to be captured during this test in % of line rate.

The test MUST provide the latency values for minimum, average and maximum, for the exact same iteration of the test.

The test MUST provide the jitter values for minimum, average and maximum, for the exact same iteration of the test.

Alternatively when a traffic generator CAN NOT be connected to all ports on the DUT, a snake test MUST be used for line rate testing, excluding latency and jitter as those became then irrelevant. The snake test consists in the following method: -connect the first and last port of the DUT to a traffic generator-connect back to back sequentially all the ports in between: port 2 to 3, port 4 to 5 etc to port n-2 to port n-1; where n is the total number of ports of the DUT-configure port 1 and 2 in the same vlan X, port 3 and 4 in the same vlan Y, etc. port n-1 and port n in the same vlan ZZZ. This snake test provides a capability to test line rate for Layer 2 and Layer 3 RFC 2544/3918 in instance where a traffic generator with only two ports is available. The latency and jitter are not to be considered with this test.

2.3 Reporting Format

The report MUST include:

-physical layer calibration information as defined into (Placeholder for definitions draft)

-number of ports used

-reading for throughput received in percentage of bandwidth, while sending 99.98% of port capacity on each port, across packet size from 64 byte all the way to 9216. As guidance, an increment of 64 byte packet size between each iteration being ideal, a 256 byte and 512 bytes being also often time used, the most common packets sizes order for the report is: 64b,128b,256b,512b,1024b,1518b,4096,8000,9216b.

The pattern for testing can be expressed using RFC 6985 [IMIX Genome: Specification of Variable Packet Sizes for Additional Testing]

-throughput needs to be expressed in % of total transmitted frames

-for packet drops, they MUST be expressed in packet count value and SHOULD be expressed in % of line rate

-for latency and jitter, values expressed in unit of time [usually microsecond or nanosecond] reading across packet size from 64 bytes to 9216 bytes

-for latency and jitter, provide minimum, average and maximum values. if different iterations are done to gather the minimum, average and maximum, it SHOULD be specified in the report along with a justification on why the information could not have been gathered at the same test iteration

-for jitter, a histogram describing the population of packets measured per latency or latency buckets is RECOMMENDED

-The tests for throughput, latency and jitter MAY be conducted as individual independent events, with proper documentation in the report but SHOULD be conducted at the same time.

3. Buffering Testing

3.1 Objective

To measure the size of the buffer of a DUT under typical|many|multiple conditions. Buffer architectures between multiple DUTs can differ and include egress buffering, shared egress buffering switch-on-chip [SoC], ingress buffering or a combination. The test methodology covers the buffer measurement regardless of buffer architecture used in the DUT.

3.2 Methodology

A traffic generator MUST be connected to all ports on the DUT.

The methodology for measuring buffering for a data-center switch is based on using known congestion of known fixed packet size along with maximum latency value measurements. The maximum latency will increase until the first packet drop occurs. At this point, the maximum latency value will remain constant. This is the point of inflexion of this maximum latency change to a constant value. There MUST be multiple ingress ports receiving known amount of frames at a known fixed size, destined for the same egress port in order to create a known congestion event. The total amount of packets sent from the oversubscribed port minus one, multiplied by the packet size represents the maximum port buffer size at the measured inflexion point.

1) Measure the highest buffer efficiency

First iteration: ingress port 1 sending line rate to egress port 2, while port 3 sending a known low amount of over subscription traffic (1% recommended) with a packet size of 64 bytes to egress port 2. Measure the buffer size value of the number of frames sent from the port sending the oversubscribed traffic up to the inflexion point multiplied by the frame size.

Second iteration: ingress port 1 sending line rate to egress port 2, while port 3 sending a known low amount of over subscription traffic (1% recommended) with same packet size 65 bytes to egress port 2. Measure the buffer size value of the number of frames sent from the port sending the oversubscribed traffic up to the inflexion point multiplied by the frame size.

Last iteration: ingress port 1 sending line rate to egress port 2, while port 3 sending a known low amount of over subscription traffic (1% recommended) with same packet size B bytes to egress port 2. Measure the buffer size value of the number of frames sent from the port sending the oversubscribed traffic up to the inflexion point multiplied by the frame size..

When the B value is found to provide the highest buffer size, this is the highest buffer efficiency

2) Measure maximum port buffer size

At fixed packet size B determined in 3.2.1, for a fixed default COS value of 0 and for unicast traffic proceed with the following:

First iteration: ingress port 1 sending line rate to egress port 2, while port 3 sending a known low amount of over subscription traffic (1% recommended) with same packet size to the egress port 2. Measure the buffer size value by multiplying the number of extra frames sent by the frame size.

Second iteration: ingress port 2 sending line rate to egress port 3, while port 4 sending a known low amount of over subscription traffic (1% recommended) with same packet size to the egress port 3. Measure the buffer size value by multiplying the number of extra frames sent by the frame size.

Last iteration: ingress port N-2 sending line rate traffic to egress port N-1, while port N sending a known low amount of over subscription traffic (1% recommended) with same packet size to the egress port N Measure the buffer size value by multiplying the number of extra frames sent by the frame size.

This test series MAY be repeated using all different COS values of traffic and then using Multicast type of traffic.

3) Measure maximum port pair buffer sizes

First iteration: ingress port 1 sending line rate to egress port 2; ingress port 3 sending line rate to egress port 4 etc. Ingress port N-1 and N will respectively over subscribe at 1% of line rate egress port 2 and port 3. Measure the buffer size value by multiplying the number of extra frames sent by the frame size for each egress port.

Second iteration: ingress port 1 sending line rate to egress port 2; ingress port 3 sending line rate to egress port 4 etc. Ingress port N-1 and N will respectively over subscribe at 1% of line rate egress port 4 and port 5. Measure the buffer size value by multiplying the number of extra frames sent by the frame size for each egress port.

Last iteration: ingress port 1 sending line rate to egress port 2; ingress port 3 sending line rate to egress port 4 etc. Ingress port N-1 and N will respectively over subscribe at 1% of line rate egress port N-3 and port N-2. Measure the buffer size value by multiplying the number of extra frames sent by the frame size for each egress port.

This test series MAY be repeated using all different COS values of traffic and then using Multicast type of traffic.

4) Measure maximum DUT buffer size with many to one ports

First iteration: ingress port 1,2,... N-1 sending each [[N-1]/[port capacity]*99.98] % of line rate per port to the N egress port.

Second iteration: ingress port 2,... N sending each [[N-1]/[port capacity]*99.98] % of line rate per port to the 1 egress port.

Last iteration: ingress port N,1,2...N-2 sending each [[N-1]/[port capacity]*99.98] % of line rate per port to the N-1 egress port.

This test series MAY be repeated using all different COS values of traffic and then using Multicast type of traffic.

Unicast traffic and then Multicast traffic SHOULD be used in order to determine the proportion of buffer for documented selection of tests. Also the COS value for the packets SHOULD be provided for each test iteration as the buffer allocation size MAY differ per COS value. It is RECOMMENDED that the ingress and egress ports are varied in a random, but documented fashion in multiple tests to measure the buffer size for each port of the DUT.

3.3 Reporting format

The report MUST include:

- The packet size used for the most efficient buffer used, along with COS value

- The maximum port buffer size for each port

- The maximum DUT buffer size

- The packet size used in the test

- The amount of over subscription if different than 1%

- The number of ingress and egress ports along with their location on the DUT.

4 Microburst Testing

4.1 Objective

To find the maximum amount of packet bursts a DUT can sustain under various configurations.

4.2 Methodology

A traffic generator MUST be connected to all ports on the DUT. In order to cause congestion, two or more ingress ports MUST bursts packets destined for the same egress port. The simplest of the setups would be two ingress ports and one egress port (2-to-1).

The burst MUST be measure with an intensity of 100%, meaning the burst of packets will be sent with a minimum inter-packet gap. The amount of packet contained in the burst will be variable and increase until there is a non-zero packet loss measured. The aggregate amount of packets from all the senders will be used to calculate the maximum amount of microburst the DUT can sustain.

It is RECOMMENDED that the ingress and egress ports are varied in multiple tests to measure the maximum microburst capacity.

The intensity of a microburst MAY be varied in order to obtain the microburst capacity at various ingress rates.

It is RECOMMENDED that all ports on the DUT will be tested simultaneously and in various configurations in order to understand all the combinations of ingress ports, egress ports and intensities.

An example would be:

First Iteration: N-1 Ingress ports sending to 1 Egress Ports

Second Iterations: N-2 Ingress ports sending to 2 Egress Ports

Last Iterations: 2 Ingress ports sending to N-2 Egress Ports

4.3 Reporting Format

The report MUST include:

- The maximum value of packets received per ingress port with the maximum burst size obtained with zero packet loss

- The packet size used in the test

- The number of ingress and egress ports along with their location on the DUT

5. Head of Line Blocking

5.1 Objective

Head-of-line blocking (HOL blocking) is a performance-limiting phenomenon that occurs when packets are held-up by the first packet ahead waiting to be transmitted to a different output port. This is defined in RFC 2889 section 5.5. Congestion Control. This section expands on RFC 2889 in the context of Data Center Benchmarking

The objective of this test is to understand the DUT behavior under head of line blocking scenario and measure the packet loss.

5.2 Methodology

In order to cause congestion, head of line blocking, groups of four ports are used. A group has 2 ingress and 2 egress ports. The first ingress port MUST have two flows configured each going to a different egress port. The second ingress port will congest the second egress port by sending line rate. The goal is to measure if there is loss for the first egress port which is not not oversubscribed.

A traffic generator MUST be connected to at least eight ports on the

DUT and SHOULD be connected using all the DUT ports.

1) Measure two groups with eight DUT ports

First iteration: measure the packet loss for two groups with consecutive ports

The first group is composed of: ingress port 1 is sending 50% of traffic to egress port 3 and ingress port 1 is sending 50% of traffic to egress port 4. Ingress port 2 is sending line rate to egress port 4. Measure the amount of traffic loss for the traffic from ingress port 1 to egress port 3.

The second group is composed of: ingress port 5 is sending 50% of traffic to egress port 7 and ingress port 5 is sending 50% of traffic to egress port 8. Ingress port 6 is sending line rate to egress port 8. Measure the amount of traffic loss for the traffic from ingress port 5 to egress port 7.

Second iteration: repeat the first iteration by shifting all the ports from N to N+1

the first group is composed of: ingress port 2 is sending 50% of traffic to egress port 4 and ingress port 2 is sending 50% of traffic to egress port 5. Ingress port 3 is sending line rate to egress port 5. Measure the amount of traffic loss for the traffic from ingress port 2 to egress port 4.

the second group is composed of: ingress port 6 is sending 50% of traffic to egress port 8 and ingress port 6 is sending 50% of traffic to egress port 9. Ingress port 7 is sending line rate to egress port 9. Measure the amount of traffic loss for the traffic from ingress port 6 to egress port 8.

Last iteration: when the first port of the first group is connected on the last DUT port and the last port of the second group is connected to the seventh port of the DUT

Measure the amount of traffic loss for the traffic from ingress port N to egress port 2 and from ingress port 4 to egress port 6.

2) Measure with N/4 groups with N DUT ports

First iteration: Expand to fully utilize all the DUT ports in increments of four. Repeat the methodology of 1) with all the group of ports possible to achieve on the device and measure for each port group the amount of traffic loss.

Second iteration: Shift by +1 the start of each consecutive ports of groups

Last iteration: Shift by N-1 the start of each consecutive ports of groups and measure the traffic loss for each port group.

5.3 Reporting Format

For each test the report MUST include:

- The port configuration including the number and location of ingress and egress ports located on the DUT

- If HOLB was observed

- Percent of traffic loss

6. Incast Stateful and Stateless Traffic

6.1 Objective

The objective of this test is to measure the effect of TCP Goodput and latency with a mix of large and small flows. The test is designed to simulate a mixed environment of stateful flows that require high rates of goodput and stateless flows that require low latency.

6.2 Methodology

In order to simulate the effects of stateless and stateful traffic on the DUT there MUST be multiple ingress ports receiving traffic destined for the same egress port. There also MAY be a mix of stateful and stateless traffic arriving on a single ingress port. The simplest setup would be 2 ingress ports receiving traffic destined to the same egress port.

One ingress port MUST be maintaining a TCP connection trough the ingress port to a receiver connected to an egress port. Traffic in the TCP stream MUST be sent at the maximum rate allowed by the traffic generator. At the same time the TCP traffic is flowing through the DUT the stateless traffic is sent destined to a receiver on the same egress port. The stateless traffic MUST be a microburst of 100% intensity.

It is RECOMMENDED that the ingress and egress ports are varied in multiple tests to measure the maximum microburst capacity.

The intensity of a microburst MAY be varied in order to obtain the microburst capacity at various ingress rates.

It is RECOMMENDED that all ports on the DUT be used in the test.

For example:

Stateful Traffic port variation:

During Iterations number of Egress ports MAY vary as well.

First Iteration: 1 Ingress port receiving stateful TCP traffic and 1 Ingress port receiving stateless traffic destined to 1 Egress Ports

Second Iteration: 2 Ingress port receiving stateful TCP traffic and 1 Ingress port receiving stateless traffic destined to 1 Egress Ports

Last Iteration: N-2 Ingress port receiving stateful TCP traffic and 1 Ingress port receiving stateless traffic destined to 1 Egress Ports

Stateless Traffic port variation:

During Iterations number of Egress ports MAY vary as well. First Iteration: 1 Ingress port receiving stateful TCP traffic and 1 Ingress port receiving stateless traffic destined to 1 Egress Ports

Second Iteration: 1 Ingress port receiving stateful TCP traffic and 2 Ingress port receiving stateless traffic destined to 1 Egress Ports

Last Iteration: 1 Ingress port receiving stateful TCP traffic and N-2 Ingress port receiving stateless traffic destined to 1 Egress Ports

6.3 Reporting Format

The report MUST include the following:

- Number of ingress and egress ports along with designation of stateful or stateless.

- TCP flow goodput

- Stateless flow latency

7. References

7.1. Normative References

[1] Bradner, S. "Benchmarking Terminology for Network Interconnection Devices", RFC 1242, July 1991.

[2] Bradner, S. and J. McQuaid, "Benchmarking Methodology for Network Interconnect Devices", RFC 2544, March 1999.

7.2. Informative References

[3] Mandeville R. and Perser J., "Benchmarking Methodology for LAN Switching Devices", RFC 2889, August 2000.

[4] Stopp D. and Hickman B., "Methodology for IP Multicast Benchmarking", BCP 26, RFC 3918, October 2004.

7.3. URL References

[5] Yanpei Chen, Rean Griffith, Junda Liu, Randy H. Katz, Anthony D. Joseph, "Understanding TCP Incast Throughput Collapse in Datacenter Networks", http://www.eecs.berkeley.edu/˜ychen2/professional/TCPIncastWREN2009.pdf"
.

Authors’ Addresses

Lucien Avramov
Cisco Systems
170 West Tasman drive
San Jose, CA 95134
United States
Phone: +1 408 526 7686
Email: lavramov@cisco.com

Jacob Rapp
Hewlett-Packard
3000 Hanover Street
Palo Alto, CA
United States
Phone: +1 650 857 3367
Email: jacob.h.rapp@hp.com