Internet Engineering Task Force L. Avramov
INTERNET-DRAFT, Intended status: Informational Google
Expires: December 23,2017 J. Rapp
June 21, 2017 VMware

Data Center Benchmarking Terminology
draft-ietf-bmwg-dcbench-terminology-18

Abstract

The purpose of this informational document is to establish definitions and describe measurement techniques for data center benchmarking, as well as it is to introduce new terminologies applicable to performance evaluations of data center network equipment. This document establishes the important concepts for benchmarking network switches and routers in the data center and, is a pre-requisite to the test methodology publication [draft-ietf-bmwg-dcbench-methodology]. Many of these terms and methods may be applicable to network equipment beyond this publication's scope as the technologies originally applied in the data center are deployed elsewhere.

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 (server to server inside the data center) in one data center and north-south (outside of the data center to server) 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. One or more of these may coexist in a single cluster and flow through a single network device simultaneously. 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 being benchmarked. These standards are good at measuring theoretical max throughput, forwarding rates and latency under testing conditions, but they do not represent real traffic patterns that may affect these networking devices. The data center networking devices covered are switches and routers.

Currently, typical data center networking devices are characterized by:

-High port density (48 ports of more)

-High speed (up to 100 GB/s currently per port)

-High throughput (line rate on all ports for Layer 2 and/or Layer 3)

-Low latency (in the microsecond or nanosecond range)

-Low amount of buffer (in the MB range per networking device)

-Layer 2 and Layer 3 forwarding capability (Layer 3 not mandatory)

The following document defines a set of definitions, metrics and terminologies including congestion scenarios, switch buffer analysis and redefines basic definitions in order to represent a wide mix of traffic conditions. The test methodologies are defined in [draft-ietf-bmwg-dcbench-methodology].

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 [RFC2119].

1.2. Definition format

Term to be defined. (e.g., Latency)

Definition: The specific definition for the term.

Discussion: A brief discussion about the term, its application and any restrictions on measurement procedures.

Measurement Units: Methodology for the measure and units used to report measurements of this term, if applicable.

2. Latency

2.1. Definition

Latency is a the amount of time it takes a frame to transit the Device Under Test (DUT). Latency is measured in units of time (seconds, milliseconds, microseconds and so on). The purpose of measuring latency is to understand the impact of adding a device in the communication path.

The Latency interval can be assessed between different combinations of events, regardless of the type of switching device (bit forwarding aka cut-through, or store-and-forward type of device). [RFC1242] defined Latency differently for each of these types of devices.

Traditionally the latency measurement definitions are:

FILO (First In Last Out)

The time interval starting when the end of the first bit of the input frame reaches the input port and ending when the last bit of the output frame is seen on the output port.

FIFO (First In First Out):

The time interval starting when the end of the first bit of the input frame reaches the input port and ending when the start of the first

bit of the output frame is seen on the output port. [RFC1242] Latency for bit forwarding devices uses these events.

LILO (Last In Last Out):

The time interval starting when the last bit of the input frame reaches the input port and the last bit of the output frame is seen on the output port.

LIFO (Last In First Out):

The time interval starting when the last bit of the input frame reaches the input port and ending when the first bit of the output frame is seen on the output port. [RFC1242] Latency for bit forwarding devices uses these events.

Another possibility to summarize the four different definitions above is to refer to the bit position as they normally occur: Input to output.

FILO is FL (First bit Last bit). FIFO is FF (First bit First bit). LILO is LL (Last bit Last bit). LIFO is LF (Last bit First bit).

This definition explained in this section in context of data center switching benchmarking is in lieu of the previous definition of Latency defined in RFC 1242, section 3.8 and is quoted here:

For store and forward devices: The time interval starting when the last bit of the input frame reaches the input port and ending when the first bit of the output frame is seen on the output port.

For bit forwarding devices: The time interval starting when the end of the first bit of the input frame reaches the input port and ending when the start of the first bit of the output frame is seen on the output port.

To accommodate both types of network devices and hybrids of the two types that have emerged, switch Latency measurements made according to this document MUST be measured with the FILO events. FILO will include the latency of the switch and the latency of the frame as well as the serialization delay. It is a picture of the 'whole' latency going through the DUT. For applications which are latency sensitive and can function with initial bytes of the frame, FIFO (or RFC 1242 Latency for bit forwarding devices) MAY be used. In all cases, the event combination used in Latency measurement MUST be reported.

2.2 Discussion

As mentioned in section 2.1, FILO is the most important measuring definition.

Not all DUTs are exclusively cut-through or store-and-forward. Data Center DUTs are frequently store-and-forward for smaller packet sizes and then adopting a cut-through behavior. The change of behavior happens at specific larger packet sizes. The value of the packet size for the behavior to change MAY be configurable depending on the DUT manufacturer. FILO covers all scenarios: Store-and-forward or cut-through. The threshold of behavior change does not matter for benchmarking since FILO covers both possible scenarios.

LIFO mechanism can be used with store forward type of switches but not with cut-through type of switches, as it will provide negative latency values for larger packet sizes because LIFO removes the serialization delay. Therefore, this mechanism MUST NOT be used when comparing latencies of two different DUTs.

2.3 Measurement Units

The measuring methods to use for benchmarking purposes are as follows:

1) FILO MUST be used as a measuring method, as this will include the latency of the packet; and today the application commonly needs to read the whole packet to process the information and take an action.

2) FIFO MAY be used for certain applications able to proceed the data as the first bits arrive, as for example for a Field-Programmable Gate Array (FPGA)

3) LIFO MUST NOT be used, because it subtracts the latency of the packet; unlike all the other methods.

3 Jitter

3.1 Definition

Jitter in the data center context is synonymous with the common term Delay variation. It is derived from multiple measurements of one-way delay, as described in RFC 3393. The mandatory definition of Delay Variation is the Packet Delay Variation (PDV) from section 4.2 of [RFC5481]. When considering a stream of packets, the delays of all packets are subtracted from the minimum delay over all packets in the stream. This facilitates assessment of the range of delay variation

(Max - Min), or a high percentile of PDV (99th percentile, for robustness against outliers).

When First-bit to Last-bit timestamps are used for Delay measurement, then Delay Variation MUST be measured using packets or frames of the same size, since the definition of latency includes the serialization time for each packet. Otherwise if using First-bit to First-bit, the size restriction does not apply.

3.2 Discussion

In addition to PDV Range and/or a high percentile of PDV, Inter-Packet Delay Variation (IPDV) as defined in section 4.1 of [RFC5481] (differences between two consecutive packets) MAY be used for the purpose of determining how packet spacing has changed during transfer, for example, to see if packet stream has become closely-spaced or "bursty". However, the Absolute Value of IPDV SHOULD NOT be used, as this collapses the "bursty" and "dispersed" sides of the IPDV distribution together.

3.3 Measurement Units

The measurement of delay variation is expressed in units of seconds. A PDV histogram MAY be provided for the population of packets measured.

4 Physical Layer Calibration

4.1 Definition

The calibration of the physical layer consists of defining and measuring the latency of the physical devices used to perform tests on the DUT.

It includes the list of all physical layer components used as listed here after:

-Type of device used to generate traffic / measure traffic

-Type of line cards used on the traffic generator

-Type of transceivers on traffic generator

-Type of transceivers on DUT

-Type of cables

-Length of cables

-Software name, and version of traffic generator and DUT

-List of enabled features on DUT MAY be provided and is recommended (especially the control plane protocols such as Link Layer Discovery Protocol, Spanning-Tree etc.). A comprehensive configuration file MAY be provided to this effect.

4.2 Discussion

Physical layer calibration is part of the end to end latency, which should be taken into acknowledgment while evaluating the DUT. Small variations of the physical components of the test may impact the latency being measured, therefore they MUST be described when presenting results.

4.3 Measurement Units

It is RECOMMENDED to use all cables of: The same type, the same length, when possible using the same vendor. It is a MUST to document the cables specifications on section 4.1 along with the test results. The test report MUST specify if the cable latency has been removed from the test measures or not. The accuracy of the traffic generator measure MUST be provided (this is usually a value in the 20ns range for current test equipment).

5 Line rate

5.1 Definition

The transmit timing, or maximum transmitted data rate is controlled by the "transmit clock" in the DUT. The receive timing (maximum ingress data rate) is derived from the transmit clock of the connected interface.

The line rate or physical layer frame rate is the maximum capacity to send frames of a specific size at the transmit clock frequency of the DUT.

The term "nominal value of Line Rate" defines the maximum speed capability for the given port; for example 1GE, 10GE, 40GE, 100GE etc.

The frequency ("clock rate") of the transmit clock in any two connected interfaces will never be precisely the same; therefore, a

tolerance is needed. This will be expressed by Parts Per Million (PPM) value. The IEEE standards allow a specific +/- variance in the transmit clock rate, and Ethernet is designed to allow for small, normal variations between the two clock rates. This results in a tolerance of the line rate value when traffic is generated from a testing equipment to a DUT.

Line rate SHOULD be measured in frames per second.

5.2 Discussion

For a transmit clock source, most Ethernet switches use "clock modules" (also called "oscillator modules") that are sealed, internally temperature-compensated, and very accurate. The output frequency of these modules is not adjustable because it is not necessary. Many test sets, however, offer a software-controlled adjustment of the transmit clock rate. These adjustments SHOULD be used to compensate the test equipment in order to not send more than the line rate of the DUT.

To allow for the minor variations typically found in the clock rate of commercially-available clock modules and other crystal-based oscillators, Ethernet standards specify the maximum transmit clock rate variation to be not more than +/- 100 PPM (parts per million) from a calculated center frequency. Therefore a DUT must be able to accept frames at a rate within +/- 100 PPM to comply with the standards.

Very few clock circuits are precisely +/- 0.0 PPM because:

1.The Ethernet standards allow a maximum of +/- 100 PPM (parts per million) variance over time. Therefore it is normal for the frequency of the oscillator circuits to experience variation over time and over a wide temperature range, among external factors.

2.The crystals, or clock modules, usually have a specific +/- PPM variance that is significantly better than +/- 100 PPM. Often times this is +/- 30 PPM or better in order to be considered a "certification instrument".

When testing an Ethernet switch throughput at "line rate", any specific switch will have a clock rate variance. If a test set is running +1 PPM faster than a switch under test, and a sustained line rate test is performed, a gradual increase in latency and eventually packet drops as buffers fill and overflow in the switch can be observed. Depending on how much clock variance there is between the two connected systems, the effect may be seen after the traffic

stream has been running for a few hundred microseconds, a few milliseconds, or seconds. The same low latency and no-packet-loss can be demonstrated by setting the test set link occupancy to slightly less than 100 percent link occupancy. Typically 99 percent link occupancy produces excellent low-latency and no packet loss. No Ethernet switch or router will have a transmit clock rate of exactly +/- 0.0 PPM. Very few (if any) test sets have a clock rate that is precisely +/- 0.0 PPM.

Test set equipment manufacturers are well-aware of the standards, and allow a software-controlled +/- 100 PPM "offset" (clock-rate adjustment) to compensate for normal variations in the clock speed of DUTs. This offset adjustment allows engineers to determine the approximate speed the connected device is operating, and verify that it is within parameters allowed by standards.

5.3 Measurement Units

"Line Rate" can be measured in terms of "Frame Rate":

Frame Rate = Transmit-Clock-Frequency / (Frame-Length*8 + Minimum_Gap + Preamble + Start-Frame Delimiter)

Minimum_Gap represents the inter frame gap. This formula "scales up" or "scales down" to represent 1 GB Ethernet, or 10 GB Ethernet and so on.

Example for 1 GB Ethernet speed with 64-byte frames: Frame Rate = 1,000,000,000 /(64*8 + 96 + 56 + 8) Frame Rate = 1,000,000,000 / 672 Frame Rate = 1,488,095.2 frames per second.

Considering the allowance of +/- 100 PPM, a switch may "legally" transmit traffic at a frame rate between 1,487,946.4 FPS and 1,488,244 FPS. Each 1 PPM variation in clock rate will translate to a 1.488 frame-per-second frame rate increase or decrease.

In a production network, it is very unlikely to see precise line rate over a very brief period. There is no observable difference between dropping packets at 99% of line rate and 100% of line rate.

Line rate can be measured at 100% of line rate with a -100PPM adjustment.

Line rate SHOULD be measured at 99,98% with 0 PPM adjustment.

The PPM adjustment SHOULD only be used for a line rate type of

measurement.

6 Buffering

6.1 Buffer

6.1.1 Definition

Buffer Size: The term buffer size represents the total amount of frame buffering memory available on a DUT. This size is expressed in B (byte); KB (kilobyte), MB (megabyte) or GB (gigabyte). When the buffer size is expressed it SHOULD be defined by a size metric stated above. When the buffer size is expressed, an indication of the frame MTU used for that measurement is also necessary as well as the cos (class of service) or dscp (differentiated services code point) value set; as often times the buffers are carved by quality of service implementation. Please refer to the buffer efficiency section for further details.

Example: Buffer Size of DUT when sending 1518 byte frames is 18 MB.

Port Buffer Size: The port buffer size is the amount of buffer for a single ingress port, egress port or combination of ingress and egress buffering location for a single port. The reason for mentioning the three locations for the port buffer is because the DUT buffering scheme can be unknown or untested, and so knowing the buffer location helps clarify the buffer architecture and consequently the total buffer size. The Port Buffer Size is an informational value that MAY be provided from the DUT vendor. It is not a value that is tested by benchmarking. Benchmarking will be done using the Maximum Port Buffer Size or Maximum Buffer Size methodology.

Maximum Port Buffer Size: In most cases, this is the same as the Port Buffer Size. In certain switch architecture called SoC (switch on chip), there is a port buffer and a shared buffer pool available for all ports. The Maximum Port Buffer Size , in terms of an SoC buffer, represents the sum of the port buffer and the maximum value of shared buffer allowed for this port, defined in terms of B (byte), KB (kilobyte), MB (megabyte), or GB (gigabyte). The Maximum Port Buffer Size needs to be expressed along with the frame MTU used for the measurement and the cos or dscp bit value set for the test.

Example: A DUT has been measured to have 3KB of port buffer for 1518 frame size packets and a total of 4.7 MB of maximum port buffer for 1518 frame size packets and a cos of 0.

Maximum DUT Buffer Size: This is the total size of Buffer a DUT can

be measured to have. It is, most likely, different than than the Maximum Port Buffer Size. It can also be different from the sum of Maximum Port Buffer Size. The Maximum Buffer Size needs to be expressed along with the frame MTU used for the measurement and along with the cos or dscp value set during the test.

Example: A DUT has been measured to have 3KB of port buffer for 1518 frame size packets and a total of 4.7 MB of maximum port buffer for 1518 B frame size packets. The DUT has a Maximum Buffer Size of 18 MB at 1500 B and a cos of 0.

Burst: The burst is a fixed number of packets sent over a percentage of linerate of a defined port speed. The amount of frames sent are evenly distributed across the interval, T. A constant, C, can be defined to provide the average time between two consecutive packets evenly spaced.

Microburst: It is a burst. A microburst is when packet drops occur when there is not sustained or noticeable congestion upon a link or device. A characterization of microburst is when the Burst is not evenly distributed over T, and is less than the constant C [C= average time between two consecutive packets evenly spaced out].

Intensity of Microburst: This is a percentage, representing the level of microburst between 1 and 100%. The higher the number the higher the microburst is. I=[1-[(TP2-Tp1)+(Tp3-Tp2)+....(TpN-Tp(n-1)] / Sum(packets)]]*100

The above definitions are not meant to comment on the ideal sizing of a buffer, rather on how to measure it. A larger buffer is not necessarily better and can cause issues with buffer bloat.

6.1.2 Discussion

When measuring buffering on a DUT, it is important to understand the behavior for each and all ports. This provides data for the total amount of buffering available on the switch. The terms of buffer efficiency here helps one understand the optimum packet size for the buffer, or the real volume of the buffer available for a specific packet size. This section does not discuss how to conduct the test methodology; instead, it explains the buffer definitions and what metrics should be provided for a comprehensive data center device buffering benchmarking.

6.1.3 Measurement Units

When Buffer is measured:

-The buffer size MUST be measured

-The port buffer size MAY be provided for each port

-The maximum port buffer size MUST be measured

-The maximum DUT buffer size MUST be measured

-The intensity of microburst MAY be mentioned when a microburst test is performed

-The cos or dscp value set during the test SHOULD be provided

6.2 Incast

6.2.1 Definition

The term Incast, very commonly utilized in the data center, refers to the traffic pattern of many-to-one or many-to-many traffic patterns. It measures the number of ingress and egress ports and the level of synchronization attributed, as defined in this section. Typically in the data center it would refer to many different ingress server ports (many), sending traffic to a common uplink (many-to-one), or multiple uplinks (many-to-many). This pattern is generalized for any network as many incoming ports sending traffic to one or few uplinks.

Synchronous arrival time: When two, or more, frames of respective sizes L1 and L2 arrive at their respective one or multiple ingress ports, and there is an overlap of the arrival time for any of the bits on the Device Under Test (DUT), then the frames L1 and L2 have a synchronous arrival times. This is called Incast regardless of in many-to-one (simpler form) or, many-to-many.

Asynchronous arrival time: Any condition not defined by synchronous arrival time.

Percentage of synchronization: This defines the level of overlap [amount of bits] between the frames L1,L2..Ln.

Example: Two 64 bytes frames, of length L1 and L2, arrive to ingress port 1 and port 2 of the DUT. There is an overlap of 6.4 bytes between the two where L1 and L2 were at the same time on the respective ingress ports. Therefore the percentage of synchronization is 10%.

Stateful type traffic defines packets exchanged with a stateful protocol such as TCP.

Stateless type traffic defines packets exchanged with a stateless protocol such as UDP.

6.2.2 Discussion

In this scenario, buffers are solicited on the DUT. In an ingress buffering mechanism, the ingress port buffers would be solicited along with Virtual Output Queues, when available; whereas in an egress buffer mechanism, the egress buffer of the one outgoing port would be used.

In either case, regardless of where the buffer memory is located on the switch architecture, the Incast creates buffer utilization.

When one or more frames having synchronous arrival times at the DUT they are considered forming an Incast.

6.2.3 Measurement Units

It is a MUST to measure the number of ingress and egress ports. It is a MUST to have a non-null percentage of synchronization, which MUST be specified.

7 Application Throughput: Data Center Goodput

7.1. Definition

In Data Center Networking, a balanced network is a function of maximal throughput and minimal loss at any given time. This is captured by the Goodput [4]. Goodput is the application-level throughput. For standard TCP applications, a very small loss can have a dramatic effect on application throughput. [RFC2647] has a definition of Goodput; the definition in this publication is a variance.

Goodput is the number of bits per unit of time forwarded to the correct destination interface of the DUT, minus any bits retransmitted.

7.2. Discussion

In data center benchmarking, the goodput is a value that SHOULD be

measured. It provides a realistic idea of the usage of the available bandwidth. A goal in data center environments is to maximize the goodput while minimizing the loss.

7.3. Measurement Units

The Goodput, G, is then measured by the following formula:

G=(S/F) x V bytes per second

-S represents the payload bytes, which does not include packet or TCP headers

-F is the frame size

-V is the speed of the media in bytes per second

Example: A TCP file transfer over HTTP protocol on a 10GB/s media.

The file cannot be transferred over Ethernet as a single continuous stream. It must be broken down into individual frames of 1500B when the standard MTU (Maximum Transmission Unit) is used. Each packet requires 20B of IP header information and 20B of TCP header information; therefore 1460B are available per packet for the file transfer. Linux based systems are further limited to 1448B as they also carry a 12B timestamp. Finally, the date is transmitted in this example over Ethernet which adds a 26B overhead per packet.

G= 1460/1526 x 10 Gbit/s which is 9.567 Gbit per second or 1.196 GB per second.

Please note: This example does not take into consideration the additional Ethernet overhead, such as the interframe gap (a minimum of 96 bit times), nor collisions (which have a variable impact, depending on the network load).

When conducting Goodput measurements please document in addition to the 4.1 section the following information:

-The TCP Stack used

-OS Versions

-NIC firmware version and model

For example, Windows TCP stacks and different Linux versions can influence TCP based tests results.

8. Security Considerations

Benchmarking activities as described in this memo are limited to technology characterization 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.

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

9. IANA Considerations

NO IANA Action is requested at this time.

10. References

10.1. Normative References

[draft-ietf-bmwg-dcbench-methodology] Avramov L. and Rapp J., "Data Center Benchmarking Methodology", RFC "draft-ietf-bmwg-dcbench-methodology", DATE (to be updated once published)

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

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

[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>

10.2. Informative References

[RFC2889] Mandeville R. and Perser J., "Benchmarking

Methodology for LAN Switching Devices", RFC 2889, August 2000, <http://www.rfc-editor.org/info/rfc2889>

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

[4] Yanpei Chen, Rean Griffith, Junda Liu, Randy H. Katz, Anthony D. Joseph, "Understanding TCP Incast Throughput Collapse in Datacenter Networks, "http://yanpeichen.com/professional/usenixLoginIncastReady.pdf"

[RFC2432] Dubray, K., "Terminology for IP Multicast Benchmarking", BCP 14, RFC 2432, DOI 10.17487/RFC2432, October 1998, <http://www.rfc-editor.org/info/rfc2432>

[RFC2647] Newman D. ,"Benchmarking Terminology for Firewall Performance" BCP 14, RFC 2647, August 1999, <http://www.rfc-editor.org/info/rfc2647>

10.3. Acknowledgments

The authors would like to thank Alfred Morton, Scott Bradner, Ian Cox, Tim Stevenson for their reviews and feedback.

Authors' Addresses

Lucien Avramov
Google
1600 Amphitheatre Parkway
Mountain View, CA 94043
United States
Phone: +1 408 774 9077
Email: lucien.avramov@gmail.com

Jacob Rapp
VMware
3401 Hillview Ave
Palo Alto, CA 94304
United States
Phone: +1 650 857 3367
Email: jrapp@vmware.com