Internet Engineering Task Force (IETF)
Request for Comments: 8238
Category: Informational
ISSN: 2070-1721

L. Avramov
Google
J. Rapp
VMware

August 2017

# Data Center Benchmarking Terminology

## Abstract

The purposes of this informational document are to establish definitions and describe measurement techniques for data center benchmarking, as well as to introduce new terminology 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 prerequisite for the test methodology document (RFC 8239). Many of these terms and methods may be applicable to network equipment beyond the scope of this document as the technologies originally applied in the data center are deployed elsewhere.

## Status of This Memo

This document is not an Internet Standards Track specification; it is published for informational purposes.

This document is a product of the Internet Engineering Task Force (IETF). It represents the consensus of the IETF community. It has received public review and has been approved for publication by the Internet Engineering Steering Group (IESG). Not all documents approved by the IESG are a candidate for any level of Internet Standard; see Section 2 of RFC 7841.

Information about the current status of this document, any errata, and how to provide feedback on it may be obtained at http://www.rfc-editor.org/info/rfc8238.

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. They can be largely east-west traffic flows (server to server inside the data center) in one data center and north-south (from the outside of the data center to the 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. All of these may coexist in a single cluster and flow through a single network device simultaneously. Benchmarking tests for 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, 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 or more).
- High speed (currently, up to 100 GB/s 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).

This document defines a set of definitions, metrics, and new terminology, including congestion scenarios and switch buffer analysis, and redefines basic definitions in order to represent a wide mix of traffic conditions. The test methodologies are defined in [RFC8239].

### 1.1. Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "NOT RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in BCP 14 [RFC2119] [RFC8174] when, and only when, they appear in all capitals, as shown here.

### 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 measurements and units used to report measurements of the term in question, if applicable.

## 2. Latency

### 2.1. Definition

Latency is the amount of time it takes a frame to transit the 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 a store-and-forward 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. Latency (as defined in [RFC1242]) 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. Latency (as defined in [RFC1242]) for store-and-forward devices uses these events.

Another possible way to summarize the four definitions above is to refer to the bit positions 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, as explained in this section in the context of data center switch benchmarking, is in lieu of the previous definition of "latency" as provided in RFC 1242, Section 3.8 and 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 that are latency sensitive and can function with initial bytes of the frame, FIFO (or, for bit-forwarding devices, latency per RFC 1242) MAY be used. In all cases, the event combinations used in latency measurements 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 change to cut-through behavior at specific larger packet sizes. The value of the packet size at which the behavior changes MAY be configurable, depending on the DUT manufacturer. FILO covers both scenarios: store-and-forward and cut-through. The threshold for the change in behavior does not matter for benchmarking, since FILO covers both possible scenarios.

The LIFO mechanism can be used with store-and-forward switches but not with cut-through 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 the 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; 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 process the data as the first bits arrive -- for example, with a Field-Programmable Gate Array (FPGA).

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

# 3. Jitter

## 3.1. Definition

In the context of the data center, jitter 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) as defined in 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 the 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 a 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 a 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. Calibration of the Physical Layer

## 4.1. Definition

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 specified here:

- Type of device used to generate traffic / measure traffic.
- Type of line cards used on the traffic generator.
- Type of transceivers on the traffic generator.
- Type of transceivers on the DUT.
- Type of cables.
- Length of cables.
- Software name and version of the traffic generator and DUT.
- A list of enabled features on the DUT MAY be provided and is recommended (especially in the case of control-plane protocols, such as the Link Layer Discovery Protocol and Spanning Tree). A comprehensive configuration file MAY be provided to this effect.

## 4.2. Discussion

Calibration of the physical layer contributes to end-to-end latency and should be taken into account when evaluating the DUT. Small variations in 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 that all cables used for testing (1) be of the same type and length and (2) come from the same vendor whenever possible. It is a MUST to document the cable specifications listed in Section 4.1, along with the test results. The test report MUST specify whether or not the cable latency has been subtracted from the test measurements. The accuracy of the traffic-generator measurements MUST be provided (for current test equipment, this is usually a value within a range of 20 ns).

# 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 (expressed as Gigabit Ethernet), 1 GE, 10 GE, 40 GE, 100 GE.

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 a 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 test equipment to a DUT.

Line rate SHOULD be measured in frames per second (FPS).

## 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 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 variance of +/- 100 PPM 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 other external factors.

2. The crystals, or clock modules, usually have a specific +/- PPM variance that is significantly better than +/- 100 PPM. Oftentimes, 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's 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 at which 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 interframe gap. This formula "scales up" or "scales down" to represent 1 GB Ethernet, 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)

= 1,000,000,000 / 672

= 1,488,095.2 FPS

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 frame-rate increase or decrease of 1.488 FPS.

In a production network, it is very unlikely that one would 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 -100 PPM adjustment.

Line rate SHOULD be measured at 99.98% with a 0 PPM adjustment.

The PPM adjustment SHOULD only be used for a line-rate 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 (bytes), KB (kilobytes), MB (megabytes), or GB (gigabytes). When the buffer size is expressed, an indication of the frame MTU (Maximum Transmission Unit) used for that measurement is also necessary, as well as the CoS (Class of Service) or DSCP (Differentiated Services Code Point) value set, as oftentimes the buffers are carved by a quality-of-service implementation. Please refer to Section 3 of [RFC8239] for further details.

Example: The Buffer Size of the 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, a single egress port, or a combination of ingress and egress buffering locations for a single port. We mention the three locations for the port buffer because the DUT’s

buffering scheme can be unknown or untested, 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 by 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 a certain type of 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 (bytes), KB (kilobytes), MB (megabytes), or GB (gigabytes). 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 3 KB of port buffer for 1518-byte frames, and a total of 4.7 MB of maximum port buffer for 1518-byte frames and a CoS of 0.

Maximum DUT Buffer Size: This is the total buffer size that a DUT can be measured to have. It is most likely different 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 3 KB of port buffer for 1518-byte frames and a total of 4.7 MB of maximum port buffer for 1518-byte frames. The DUT has a Maximum Buffer Size of 18 MB at 1500 B and a CoS of 0.

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

Microburst: A microburst is a type of burst where packet drops occur when there is not sustained or noticeable congestion on a link or device. One characteristic of a microburst is when the burst is not evenly distributed over T and is less than the constant C (C = the average time between two evenly spaced consecutive packets).

Intensity of Microburst: This is a percentage and represents the level, between 1 and 100%, of the microburst. 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 but rather on how to measure it. A larger buffer is not necessarily better and can cause issues with bufferbloat.

### 6.1.2. Discussion

When measuring buffering on a DUT, it is important to understand the behavior of each and every port. This provides data for the total amount of buffering available on the switch. The terms of buffer efficiency help 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 comprehensive data center device-buffering benchmarking.

### 6.1.3. Measurement Units

When the DUT 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 the 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 many-to-one or many-to-many traffic patterns. As defined in this section, it measures the number of ingress and egress ports and the percentage of synchronization attributed to them. 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 a few uplinks.

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

Asynchronous arrival time: This is any condition not defined by "synchronous arrival time".

Percentage of synchronization: This defines the level of overlap (amount of bits) between frames of sizes L1,L2..Ln.

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

Stateful traffic: Stateful traffic is packets exchanged with a stateful protocol, such as TCP.

Stateless traffic: Stateless traffic is packets exchanged with a stateless protocol, such as UDP.

### 6.2.2. Discussion

In this scenario, buffers are used on the DUT. In an ingress buffering mechanism, the ingress port buffers would be used along with virtual output queues, when available, whereas in an egress buffering mechanism, the egress buffer of the one outgoing port would be used.

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

When one or more frames have synchronous arrival times at the DUT, they are considered to be 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 [TCP-INCAST]. Goodput is the application-level throughput. For standard TCP applications, a very small loss can have a dramatic effect on application throughput. [RFC2647] provides a definition of Goodput; the definition in this document is a variant of that definition.

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 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, not including 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 on 10 GB/s media.

The file cannot be transferred over Ethernet as a single continuous stream. It must be broken down into individual frames of 1500 B when the standard MTU is used. Each packet requires 20 B of IP header information and 20 B of TCP header information; therefore, 1460 B are available per packet for the file transfer. Linux-based systems are further limited to 1448 B, as they also carry a 12 B timestamp. Finally, in this example the date is transmitted over Ethernet, which adds 26 B of overhead per packet to 1500 B, increasing it to 1526 B.

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

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 does it account for collisions (which have a variable impact, depending on the network load).

When conducting Goodput measurements, please document, in addition to the items listed in Section 4.1, the following information:

- The TCP stack used.
- OS versions.
- Network Interface Card (NIC) firmware version and model.

For example, Windows TCP stacks and different Linux versions can influence TCP-based test 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

This document does not require any IANA actions.

## 10. References

### 10.1. Normative References

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

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

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

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

[RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017, <https://www.rfc-editor.org/info/rfc8174>.

[RFC8239] Avramov, L. and J. Rapp, "Data Center Benchmarking Methodology", RFC 8239, DOI 10.17487/RFC8239, August 2017, <https://www.rfc-editor.org/info/rfc8239>.

## 10.2. Informative References

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

[RFC2647] Newman, D., "Benchmarking Terminology for Firewall Performance", RFC 2647, DOI 10.17487/RFC2647, August 1999, <https://www.rfc-editor.org/info/rfc2647>.

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

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

[TCP-INCAST] Chen, Y., Griffith, R., Zats, D., Joseph, A., and R. Katz, "Understanding TCP Incast and Its Implications for Big Data Workloads", April 2012, <http://yanpeichen.com/professional/usenixLoginIncastReady.pdf>.

## Acknowledgments

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

## Authors' Addresses

Lucien Avramov
Google
1600 Amphitheatre Parkway
Mountain View, CA 94043
United States of America

Email: lucien.avramov@gmail.com

Jacob Rapp
VMware
3401 Hillview Ave.
Palo Alto, CA 94304
United States of America

Email: jhrapp@gmail.com