Network Working Group
Internet-Draft
Updates: 2544 (if approved)
Intended status: Informational
Expires: June 21, 2021

A. Morton
AT&T Labs

December 18, 2020

Updates for the Back-to-back Frame Benchmark in RFC 2544
draft-ietf-bmwg-b2b-frame-04

Abstract

Fundamental Benchmarking Methodologies for Network Interconnect Devices of interest to the IETF are defined in RFC 2544. This memo updates the procedures of the test to measure the Back-to-back frames Benchmark of RFC 2544, based on further experience.

This memo updates Section 26.4 of RFC 2544.

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.

Status of This Memo

Table of Contents

1. Introduction

The IETF's fundamental Benchmarking Methodologies are defined in [RFC2544], supported by the terms and definitions in [RFC1242], and [RFC2544] actually obsoletes an earlier specification, [RFC1944]. Over time, the benchmarking community has updated [RFC2544] several times, including the Device Reset Benchmark [RFC6201], and the important Applicability Statement [RFC6815] concerning use outside the Isolated Test Environment (ITE) required for accurate benchmarking. Other specifications implicitly update [RFC2544], such as the IPv6 Benchmarking Methodologies in [RFC5180].

Recent testing experience with the Back-to-back Frame test and Benchmark in Section 26.4 of [RFC2544] indicates that an update is warranted [OPNFV-2017] [VSPERF-b2b]. In particular, analysis of the results indicates that buffer size matters when compensating for interruptions of software packet processing, and this finding increases the importance of the Back-to-back frame characterization described here. This memo describes additional rationale and provides the updated method.

[RFC2544] (which obsoletes [RFC1944]) provides its own Requirements Language consistent with [RFC2119], since [RFC1944] pre-dates [RFC2119] and all three memos share common authorship. Today,[RFC8174] clarifies the usage of Requirements Language, so the requirements presented in this memo are expressed in [RFC8174] terms, and intended for those performing/reporting laboratory tests to improve clarity and repeatability, and for those designing devices that facilitate these tests.

2. Scope and Goals

The scope of this memo is to define an updated method to unambiguously perform tests, measure the benchmark(s), and report the results for Back-to-back Frames (presently described in Section 26.4 of [RFC2544]).

The goal is to provide more efficient test procedures where possible, and to expand reporting with additional interpretation of the results. The tests described in this memo address the cases in which the maximum frame rate of a single ingress port cannot be transferred loss-free to an egress port (for some frame sizes of interest).

[RFC2544] Benchmarks rely on test conditions with constant frame sizes, with the goal of understanding what network device capability has been tested. Tests with the smallest size stress the header processing capacity, and tests with the largest size stress the overall bit processing capacity. Tests with sizes in-between may determine the transition between these two capacities. However, conditions simultaneously sending a mixture of Internet frame sizes (IMIX), such as those described in [RFC6985], MUST NOT be used in Back-to-back Frame testing.

Section 3 of [RFC8239] describes buffer size testing for physical networking devices in a data center. The [RFC8239] methods measure buffer latency directly with traffic on multiple ingress ports that overload an egress port on the Device Under Test (DUT) and are not subject to the revised calculations presented in this memo. Likewise, the methods of [RFC8239] SHOULD be used for test cases where the egress port buffer is the known point of overload.

3. Motivation

Section 3.1 of [RFC1242] describes the rationale for the Back-to-back Frames Benchmark. To summarize, there are several reasons that devices on a network produce bursts of frames at the minimum allowed spacing; and it is, therefore, worthwhile to understand the Device Under Test (DUT) limit on the length of such bursts in practice. Also, [RFC1242] states:

> "Tests of this parameter are intended to determine the extent of data buffering in the device."

After this test was defined, there have been occasional discussions of the stability and repeatability of the results, both over time and across labs. Fortunately, the Open Platform for Network Function Virtualization (OPNFV) VSPERF project's Continuous Integration (CI) [VSPERF-CI] testing routinely repeats Back-to-back Frame tests to verify that test functionality has been maintained through development of the test control programs. These tests were used as a basis to evaluate stability and repeatability, even across lab set-ups when the test platform was migrated to new DUT hardware at the end of 2016.

When the VSPERF CI results were examined [VSPERF-b2b], several aspects of the results were considered notable:

1. Back-to-back Frame Benchmark was very consistent for some fixed frame sizes, and somewhat variable for other frame sizes.

2. The number of Back-to-back Frames with zero loss reported for large frame sizes was unexpectedly long (translating to 30 seconds of buffer time), and no explanation or measurement limit condition was indicated. It was important that the buffering time calculations were part of the referenced testing and analysis[VSPERF-b2b], because the calculated buffer times of 30 seconds for some frame sizes were clearly wrong or highly suspect. On the other hand, a result expressed only as a large number of Back-to-back Frames does not permit such an easy comparison with reality.

3. Calculation of the extent of buffer time in the DUT helped to explain the results observed with all frame sizes (for example, tests with some frame sizes cannot exceed the frame header processing rate of the DUT and thus no buffering occurs; therefore, the results depended on the test equipment and not the DUT).

4. It was found that a better estimate of the DUT buffer time could be calculated using measurements of both the longest burst in frames without loss and results from the Throughput tests conducted according to Section 26.1 of [RFC2544]. It is apparent that the DUT's frame processing rate empties the buffer during a trial and tends to increase the "implied" buffer size estimate (measured according to Section 26.4 of [RFC2544] because many frames have departed the buffer when the burst of frames ends). A calculation using the Throughput measurement can reveal a "corrected" buffer size estimate.

Further, if the Throughput tests of Section 26.1 of [RFC2544] are conducted as a prerequisite test, the number of frame sizes required for Back-to-back Frame Benchmarking can be reduced to one or more of the small frame sizes, or the results for large frame sizes can be noted as invalid in the results if tested anyway (these are the larger frame sizes for which the back-to-back frame rate cannot exceed the frame header processing rate of the DUT and little or no buffering occurs).

The material below provides the details of the calculation to estimate the actual buffer storage available in the DUT, using results from the Throughput tests for each frame size, and the maximum theoretical frame rate for the DUT links (which constrain the minimum frame spacing).

In reality, there are many buffers and packet header processing steps in a typical DUT. The simplified model used in these calculations for the DUT includes a packet header processing function with limited rate of operation, as shown below:

```
                     |------------ DUT --------|
Generator -> Ingress -> Buffer -> HeaderProc -> Egress -> Receiver
```

So, in the Back-to-back Frame testing:

1. The ingress burst arrives at Max Theoretical Frame Rate, and initially the frames are buffered.

2. The packet header processing function (HeaderProc) operates at the "Measured Throughput" (Section 26.1 of [RFC2544]), removing frames from the buffer (this is the best approximation we have).

3. Frames that have been processed are clearly not in the buffer, so the Corrected DUT buffer time equation (Section 5.4) estimates and removes the frames that the DUT forwarded on egress during the burst. We define buffer time as the number of frames

occupying the buffer divided by the Maximum Theoretical Frame Rate (on ingress) for the frame size under test.

4. A helpful concept is the buffer filling rate, which is the difference between the Max Theoretical Frame Rate (ingress) and the Measured Throughput (HeaderProc on egress). If the actual buffer size in frames was known, the time to fill the buffer during a measurement can be calculated using the filling rate as a check on measurements. However, the buffer in the model represents many buffers of different sizes in the DUT data path.

Knowledge of approximate buffer storage size (in time or bytes) may be useful to estimate whether frame losses will occur if DUT forwarding is temporarily suspended in a production deployment, due to an unexpected interruption of frame processing (an interruption of duration greater than the estimated buffer would certainly cause lost frames). In Section 5, the calculations for the correct buffer time use the combination of offered load at Max Theoretical Frame Rate and header processing speed at 100% of Measured Throughput. Other combinations are possible, such as changing the percent of measured Throughput to account for other processes reducing the header processing rate.

The presentation of OPNFV VSPERF evaluation and development of enhanced search algorithms [VSPERF-BSLV] was discussed at IETF-102. The enhancements are intended to compensate for transient interrupts that may cause loss at near-Throughput levels of offered load. Subsequent analysis of the results indicates that buffers within the DUT can compensate for some interrupts, and this finding increases the importance of the Back-to-back frame characterization described here.

4. Prerequisites

The Test Setup MUST be consistent with Figure 1 of [RFC2544], or Figure 2 when the tester's sender and receiver are different devices. Other mandatory testing aspects described in [RFC2544] MUST be included, unless explicitly modified in the next section.

The ingress and egress link speeds and link layer protocols MUST be specified and used to compute the maximum theoretical frame rate when respecting the minimum inter-frame gap.

The test results for the Throughput Benchmark conducted according to Section 26.1 of [RFC2544] for all [RFC2544]-RECOMMENDED frame sizes MUST be available to reduce the tested frame size list, or to note invalid results for individual frame sizes (because the burst length may be essentially infinite for large frame sizes).

Note that:

- o the Throughput and the Back-to-back Frame measurement configuration traffic characteristics (unidirectional or bi-directional, and number of flows generated) MUST match.

- o the Throughput measurement MUST be under zero-loss conditions, according to Section 26.1 of [RFC2544].

The Back-to-back Benchmark described in Section 3.1 of [RFC1242] MUST be measured directly by the tester, where buffer size is inferred from Back-to-back Frame bursts and associated packet loss measurements. Therefore, sources of packet loss that are unrelated to consistent evaluation of buffer size SHOULD be identified and removed or mitigated. Example sources include:

- o On-path active components that are external to the DUT

- o Operating system environment interrupting DUT operation

- o Shared resource contention between the DUT and other off-path component(s) impacting DUT's behaviour, sometimes called the "noisy neighbour" problem with virtualized network functions.

Mitigations applicable to some of the sources above are discussed in Section 5.2, with the other measurement requirements described below in Section 5.

5. Back-to-back Frames

Objective: To characterize the ability of a DUT to process back-to-back frames as defined in [RFC1242].

The Procedure follows.

5.1. Preparing the list of Frame sizes

From the list of RECOMMENDED frame sizes (Section 9 of [RFC2544]), select the subset of frame sizes whose measured Throughput (during prerequisite testing) was less than the Maximum Theoretical Frame Rate of the DUT/test-set-up. These are the only frame sizes where it is possible to produce a burst of frames that cause the DUT buffers to fill and eventually overflow, producing one or more discarded frames.

5.2. Test for a Single Frame Size

Each trial in the test requires the tester to send a burst of frames (after idle time) with the minimum inter-frame gap, and to count the corresponding frames forwarded by the DUT.

The duration of the trial includes three REQUIRED components:

1. The time to send the burst of frames (at the back-to-back rate), determined by the search algorithm.

2. The time to receive the transferred burst of frames (at the [RFC2544] Throughput rate), possibly truncated by buffer overflow, and certainly including the latency of the DUT.

3. At least 2 seconds not overlapping the time to receive the burst (2.), to ensure that DUT buffers have depleted. Longer times MUST be used when conditions warrant, such as when buffer times >2 seconds are measured or when burst sending times are >2 seconds, but care is needed since this time component directly increases trial duration and many trials and tests comprise a complete benchmarking study.

The upper search limit for the time to send each burst MUST be configurable, to values as high as 30 seconds (buffer time results reported at or near the configured upper limit are likely invalid, and the test MUST be repeated with a higher search limit).

If all frames have been received, the tester increases the length of the burst according to the search algorithm and performs another trial.

If the received frame count is less than the number of frames in the burst, then the limit of DUT processing and buffering may have been exceeded, and the burst length is determined by the search algorithm for the next trial (the burst length is typically reduced, but see below).

Classic search algorithms have been adapted for use in benchmarking, where the search requires discovery of a pair of outcomes, one with no loss and another with loss, at load conditions within the acceptable tolerance or accuracy. Conditions encountered when benchmarking the Infrastructure for Network Function Virtualization require algorithm enhancement. Fortunately, the adaptation of Binary Search, and an enhanced Binary Search with Loss Verification have been specified in clause 12.3 of [TST009]. These algorithms can easily be used for Back-to-back Frame benchmarking by replacing the Offered Load level with burst length in frames. [TST009] Annex B

describes the theory behind the enhanced Binary Search with Loss Verification algorithm.

There is also promising work-in-progress that may prove useful in Back-to-back Frame benchmarking. [I-D.vpolak-mkonstan-bmwg-mlrsearch] and [I-D.vpolak-bmwg-plrsearch] are two such examples.

Either the [TST009] Binary Search or Binary Search with Loss Verification algorithms MUST be used, and input parameters to the algorithm(s) MUST be reported.

The tester usually imposes a (configurable) minimum step size for burst length, and the step size MUST be reported with the results (as this influences the accuracy and variation of test results).

The original Section 26.4 of [RFC2544] definition is stated below:

> The Back-to-back Frame value is the longest burst of frames that the DUT can successfully process and buffer without frame loss, as determined from the series of trials.

5.3. Test Repetition and Benchmark

On this topic, Section 26.4 of [RFC2544] requires:

> The trial length MUST be at least 2 seconds and SHOULD be repeated at least 50 times with the average of the recorded values being reported.

Therefore, the Back-to-back Frame Benchmark is the average of burst length values over repeated tests to determine the longest burst of frames that the DUT can successfully process and buffer without frame loss. Each of the repeated tests completes an independent search process.

In this update, the test MUST be repeated N times (the number of repetitions is now a variable that must be reported),for each frame size in the subset list, and each Back-to-back Frame value made available for further processing (below).

5.4. Benchmark Calculations

For each frame size, calculate the following summary statistics for longest Back-to-back Frame values over the N tests:

o Average (Benchmark)

o Minimum

o Maximum

o Standard Deviation

Further, calculate the Implied DUT Buffer Time and the Corrected DUT Buffer Time in seconds, as follows:

Implied DUT Buffer Time =

Average num of Back-to-back Frames / Max Theoretical Frame Rate

The formula above is simply expressing the burst of frames in units of time.

The next step is to apply a correction factor that accounts for the DUT's frame forwarding operation during the test (assuming the simple model of the DUT composed of a buffer and a forwarding function, described in Section 3).

Corrected DUT Buffer Time =

```
                 /                                         \
   Implied DUT    |Implied DUT        Measured Throughput    |
=  Buffer Time -  |Buffer Time * -------------------------- |
                  |               Max Theoretical Frame Rate |
                 \                                         /
```

where:

1. The "Measured Throughput" is the [RFC2544] Throughput Benchmark for the frame size tested, as augmented by methods including the Binary Search with Loss Verification algorithm in [TST009] where applicable, and MUST be expressed in frames per second in this equation.

2. The "Max Theoretical Frame Rate" is a calculated value for the interface speed and link layer technology used, and MUST be expressed in frames per second in this equation.

The term on the far right in the formula for Corrected DUT Buffer Time accounts for all the frames in the Burst that were transmitted by the DUT *while the Burst of frames were sent in*. So, these frames are not in the buffer and the buffer size is more accurately estimated by excluding them.

6. Reporting

The back-to-back frame results SHOULD be reported in the format of a table with a row for each of the tested frame sizes. There SHOULD be columns for the frame size and for the resultant average frame count for each type of data stream tested.

The number of tests Averaged for the Benchmark, N, MUST be reported.

The Minimum, Maximum, and Standard Deviation across all complete tests SHOULD also be reported (they are referred to as "Min,Max,StdDev" in the table below).

The Corrected DUT Buffer Time SHOULD also be reported.

If the tester operates using a limited maximum burst length in frames, then this maximum length SHOULD be reported.

Frame Size, octets	Ave B2B Length, frames	Min,Max,StdDev	Corrected Buff Time, Sec
64	26000	25500,27000,20	0.00004

Back-to-Back Frame Results

Static and configuration parameters (reported with the table above):

Number of test repetitions, N

Minimum Step Size (during searches), in frames.

If the tester has a specific (actual) frame rate of interest (less than the Throughput rate), it is useful to estimate the buffer time at that actual frame rate:

```
Actual Buffer Time =
                                   Max Theoretical Frame Rate
     = Corrected DUT Buffer Time * --------------------------
                                       Actual Frame Rate
```

and report this value, properly labeled.

7. 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 other constraints of[RFC2544].

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. See [RFC6815].

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

The DUT developers are commonly independent from the personnel and institutions conducting benchmarking studies. DUT developers might have incentives to alter the performance of the DUT if the test conditions can be detected. Special capabilities SHOULD NOT exist in the DUT/SUT specifically for benchmarking purposes. Procedures described in this document are not designed to detect such activity. Additional testing outside of the scope of this document would be needed and has been used successfully in the past to discover such malpractices.

Any implications for network security arising from the DUT/SUT SHOULD be identical in the lab and in production networks.

8. IANA Considerations

This memo makes no requests of IANA.

9. Acknowledgements

Thanks to Trevor Cooper, Sridhar Rao, and Martin Klozik of the VSPERF project for many contributions to the early testing [VSPERF-b2b]. Yoshiaki Itou has also investigated the topic, and made useful suggestions. Maciek Konstantyowicz and Vratko Polak also provided many comments and suggestions based on extensive integration testing and resulting search algorithm proposals - the most up-to-date feedback possible. Tim Carlin also provided comments and support for the draft. Warren Kumari's review improved readability in several key passages. David Black, Martin Duke, and Scott Bradner's comments improved the clarity and configuration advice on trial duration. Malisa Vucinic suggested additional text on DUT design cautions in the Security Considerations section.

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>.

[RFC6985] Morton, A., "IMIX Genome: Specification of Variable Packet Sizes for Additional Testing", RFC 6985, DOI 10.17487/RFC6985, July 2013, <https://www.rfc-editor.org/info/rfc6985>.

[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>.

[TST009] Morton, A., "ETSI GS NFV-TST 009 V3.4.1 (2020-12), "Network Functions Virtualisation (NFV) Release 3; Testing; Specification of Networking Benchmarks and Measurement Methods for NFVI"", December 2020, <https://www.etsi.org/deliver/etsi_gs/NFV-TST/001_099/009/03.04.01_60/gs_NFV-TST009v030401p.pdf>.

10.2. Informative References

[I-D.vpolak-bmwg-plrsearch]
Konstantynowicz, M. and V. Polak, "Probabilistic Loss Ratio Search for Packet Throughput (PLRsearch)", draft-vpolak-bmwg-plrsearch-03 (work in progress), March 2020.

[I-D.vpolak-mkonstan-bmwg-mlrsearch]
Konstantynowicz, M. and V. Polak, "Multiple Loss Ratio Search for Packet Throughput (MLRsearch)", draft-vpolak-mkonstan-bmwg-mlrsearch-03 (work in progress), March 2020.

[OPNFV-2017]
Cooper, T., Morton, A., and S. Rao, "Dataplane Performance, Capacity, and Benchmarking in OPNFV", June 2017, <https://wiki.opnfv.org/download/attachments/10293193/VSPERF-Dataplane-Perf-Cap-Bench.pptx?api=v2>.

[RFC1944] Bradner, S. and J. McQuaid, "Benchmarking Methodology for Network Interconnect Devices", RFC 1944, DOI 10.17487/RFC1944, May 1996, <https://www.rfc-editor.org/info/rfc1944>.

[RFC5180] Popoviciu, C., Hamza, A., Van de Velde, G., and D. Dugatkin, "IPv6 Benchmarking Methodology for Network Interconnect Devices", RFC 5180, DOI 10.17487/RFC5180, May 2008, <https://www.rfc-editor.org/info/rfc5180>.

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

[RFC6815] Bradner, S., Dubray, K., McQuaid, J., and A. Morton, "Applicability Statement for RFC 2544: Use on Production Networks Considered Harmful", RFC 6815, DOI 10.17487/RFC6815, November 2012, <https://www.rfc-editor.org/info/rfc6815>.

[VSPERF-b2b]
Morton, A., "Back2Back Testing Time Series (from CI)", June 2017, <https://wiki.opnfv.org/display/vsperf/Traffic+Generator+Testing#TrafficGeneratorTesting-AppendixB:Back2BackTestingTimeSeries(fromCI)>.

[VSPERF-BSLV]
Morton, A. and S. Rao, "Evolution of Repeatability in Benchmarking: Fraser Plugfest (Summary for IETF BMWG)", July 2018, <https://datatracker.ietf.org/meeting/102/materials/slides-102-bmwg-evolution-of-repeatability-in-benchmarking-fraser-plugfest-summary-for-ietf-bmwg-00>.

[VSPERF-CI]
Tahhan, M., "OPNFV VSPERF CI", June 2019,
<https://wiki.opnfv.org/display/vsperf/VSPERF+CI>.

Author's Address

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