Bridge Out: Benchmarking Methodology Extensions for Data Center Bridging
                              Devices
                 draft-player-dcb-benchmarking-03.txt

Abstract

   Existing benchmarking methodologies are based on the assumption that
   networking devices will impartially drop network traffic at their
   performance limits. Data Center Bridging (DCB) devices, however,
   will attempt to throttle prioritized traffic from network endpoints
   before those limits are reached in order to minimize the probability
   of frame loss for high value traffic. Hence, existing methodologies
   based around indiscriminate frame loss are inappropriate for DCB
   devices. This document takes the basic benchmarking ideas based on
   loss and extends them to support "lossless" Ethernet devices.

Status of this Memo

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

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

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

   This Internet-Draft will expire on April 22, 2011.

Copyright Notice

Table of Contents

1. Introduction

   This document is intended to provide a methodology for benchmarking
   Data Center Bridging (DCB) devices that support Priority-based Flow
   Control (PFC). It extends the methodologies already defined in
   [RFC2544] and [RFC2889].

   This memo primarily deals with devices which use Priority-based Flow
   Control, as defined in IEEE specification 802.1Qbb, to actively
   manage the transmission rate of multiple classes of traffic in order
   to minimize forwarding delay and frame loss for high priority
   traffic.

2. Requirements

   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 [RFC2119].

3. Terminology

   As the terminology used by [RFC4689] is specific to IP layer testing,
   a number of existing terms require clarification when used in the DCB
   benchmarking context. Additionally, a number of new terms are also
   presented to clarify concepts not clearly defined within the scope of
   [RFC4689].

      Classification: As stated in [RFC4689], Classification is the
      selection of packets according to defined rules. In the context
      of DCB benchmarking, the Classification criterion is the value of
      the 802.1p priority code point field in the 802.1Q VLAN header of
      an Ethernet frame.

      Classification Group: A collection of traffic streams that belong
      to a single Classification. A Conformance Vector MAY be
      associated with a Classification Group.

      Classification Profile: The set of all Classification Groups
      involved in a benchmarking test.

      Conformance Vector: A set of measurable stream result bounds, e.g.
      latency, jitter, sequencing, etc., that specify whether a frame is
      Conformant or Non-conformant. Conformance vectors are optional
      for all DCB benchmarking tests.

Congestion Management: In the context of DCB benchmarking, Congestion Management occurs when the DUT/SUT transmits Priority-based Flow Control (PFC) Pause frames.

Forwarding Congestion: In the context of DCB benchmarking, Forwarding Congestion is extended to include the observation of PFC pause frame transmissions from the DUT.

Intended Load: In this document, the Intended Load refers to the summation of the Intended Vectors for all Classification Groups.

Offered Load: In this document, the Offered Load refers to the summation of the Offered Vectors for all Classification Groups.

Queue Congestion: Queue congestion occurs when a DUT/SUT uses Congestion Management on a set of traffic Classifications. The congestion Classifications correspond to the congested queues in the DUT/SUT.

Queueput: The maximum Offered Load than can be transmitted into a DUT/SUT such that every transmitted frame matches a specific Classification rule, the DUT/SUT does NOT use priority-based flow control mechanisms to manage the ingress traffic rate of the Classification(s) of interest, and all ingress frames are forwarded to the correct egress port. A DUT may have a different Queueput value for each configured Classification.

XOFF Frame: A Priority-based flow control pause frame that instructs the DUT to pause one or more VLAN priorities.

XON Frame: A Priority-based flow control pause frame that instructs the DUT to resume transmission on one or more VLAN priorities.

[4](#). General Considerations

[4.1](#). Classifications

Data Center Bridging devices SHOULD be tested with multiple Classifications. Testing with a single Classification provides no means to test and measure a device's ability to differentiate forwarding behavior for different traffic classes.

[4.2](#). Congestion

For devices capable of forwarding traffic at line rate, explicit congestion MUST be created via the test tool to benchmark queue

performance. Possible methods for accomplishing this on a DUT with n
ports include, but are not limited to:

1. Test full-mesh traffic patterns on (n-1) ports while using 1 port
    as a multicast transmitter with (n-1) multicast receivers.

2. Test full-mesh traffic patterns on (n-1) ports while generating
    partially meshed traffic between 1 and (n-1) ports.

3. Use partially meshed traffic patterns with x ports transmitting
    to y ports where x > y and x + y = n.

4.3. Test Traffic

The lock-step traffic pattern, as described in section 5.1.3 of
[RFC2889], is specifically NOT required for DCB testing for two
reasons:

1. Such patterns are not meaningful for high speed Ethernet devices
    due to the transmission clock variance allowed by the IEEE 802.3
    Ethernet specification.

2. Flow control mechanisms would quickly break such patterns when
    activated.

4.4. Tester Capabilities

4.4.1. Frame Formats

This testing document does not mandate the use of any particular
frame format for testing. Any frame that can be legally forwarded by
the DUT/SUT MAY be used provided that the test instrument can make
the following distinctions for each frame:

1. The test tool MUST be able to distinguish test frames from non-
    test frames.

2. The test tool MUST be able to determine whether each test frame
    is forwarded to the correct egress port.

3. The test tool MUST be able to determine whether each received
    frame conforms or does not conform to the Conformance Vector of
    the frame's Classification Group, if applicable.

4.4.2. Pause Response Time

   To accurately measure the performance of a Priority-based Flow
   Control capable DUT, the test tool MUST be able to respond to PFC
   pause frames. Additionally, the test tool MUST respond to all
   received pause frames in the time period specified in the IEEE
   802.1Qbb specification.

5. Test Setup

   This document extends the general test setup described in section 3
   of [RFC2889] and section 6 of [RFC2544] to the benchmarking of Data
   Center Ethernet switching devices. [RFC2889] and [RFC2544] describe
   benchmarking methodologies for networking devices that intentionally
   drop frames at their performance limits. In DCB networks, the DUT
   will transmit PFC Pause frames as a Congestion Management method to
   throttle network endpoints, thus minimizing the probability of frame
   loss in the network.

5.1. Test Traffic

5.1.1. Traffic Classification

   Since DCB devices are expected to support multiple traffic
   Classifications, it is RECOMMENDED to benchmark DCB devices with
   multiple Classification Groups.

5.1.2. Trial Duration

   The RECOMMENDED trial duration is 300 seconds. However other
   durations MAY be used. Additionally, a running trial MAY be aborted
   once the test tool determines that the currently running trial has
   failed, e.g. QoS bounds exceeded, packet loss detected on a lossless
   queue, etc.

5.1.3. Frame Measurements

   Packet Conformance MUST be determined for all test frames on a per
   frame basis. The method specified for measuring Latency in
   [RFC2544], e.g. measuring the latency of a single test frame in a
   traffic flow, is unsuitable for DCB benchmarking.

5.1.3.1. Forwarding Delay and Latency

   Multiple methods exist for measuring the time it takes a test frame
   to be forwarded by a DUT. However, both of the methods discussed in
   [RFC1242] are unsuitable for testing DCB devices, as many DCB devices

alternate between both "store and forward" and "bit forwarding" behavior depending upon their queue congestion. Hence, the only RECOMMENDED method for measuring the time it takes a DUT to forward a test frame is "Forwarding Delay" as described in [RFC4689].

5.1.4. Frame Sizes

5.1.4.1. Ethernet

The recommended frame sizes for Ethernet testing are 64, 128, 256, 512, 1024, 1280, 1518, 4096, 8192, and 9216 as per [RFC5180]. Note that these frame sizes include the Ethernet CRC and VLAN header.

5.1.4.1.1. Fiber Channel over Ethernet

FCoE test traffic introduces a number of frame size constraints that make the default frame sizes specified in [RFC5180] unusable:

1. FCoE frames contain an encapsulated Fiber Channel frame. Due to the method of encapsulation used, all FCoE frames MUST be a multiple of 4 bytes. See [RFC3643].

2. Test tools may need to include a test payload in addition to the encapsulated Fiber Channel frame to meet the requirements specified in Section 4.4.1.

3. The maximum supported frame size for FCoE is 2176 bytes.

Due to these constraints, the recommended frame sizes for FCoE testing are 128, 256, 512, 1024, 1280, 1520, 2176, and the smallest FCoE frame size supported by the test tool. Note that these frame sizes include both the Ethernet CRC and VLAN header.

5.1.5. Burst Sizes

As per [RFC2285], the burst size specifies the number of test frames in a burst. To simulate bursty traffic, the test tool MAY send a burst of test traffic with the minimum, legal Inter-Frame Gap (IFG) between frames in the burst followed by a larger Inter-Burst Gap (IBG) between sequential bursts. Note that burst sizes are only applicable to test traffic when the Offered Load of the test ports is less than the Maximum Offered Load (MOL) of those ports. Additionally, a burst size of 1 specifies a constant load, e.g. non-bursty traffic.

6. Benchmarking Tests

6.1. Pause Response Time

6.1.1. Objective

   To determine the amount of time required for the DUT to respond to
   priority-based flow control pause frames.

6.1.2. Setup Parameters

   The following parameters MUST be defined. Each variable is
   configured with the following considerations.

      Each Classification Group MUST be listed. For each classification
      group, the following parameters MUST be specified:

         Codepoint - For DCB tests, the codepoint is the VLAN priority.

         Frame Size - The frame size includes both the CRC and VLAN
         header. See Section 5.1.4 for recommended frame sizes.

         Burst Size - The burst size specifies the number of frames
         transmitted with the minimum legal IFG before pausing. See
         Section 5.1.5.

         Intended Vector - The intended vector SHOULD specify the
         intended rate of test traffic specified as a percentage of port
         load.

         Traffic Pattern - The traffic distribution and traffic
         orientation used for this Classification.

         Conformance Vector - The conformance vector is optional, but
         MUST be defined if used.

      Priority-based Flow Control - PFC mechanisms MUST be enabled.

      Background Traffic - Background traffic MAY be present.

      PFC Pause Parameters:

         Queue(s) - A list of one or more VLAN priorities the test tool
         should attempt to pause.

         Pause Value - The quanta value to use in the XOFF frame(s).

XON Delay - The amount of time to pause the DUT before sending
a XON frame. Note that if the XON Delay is larger than the
Pause Value, the test tool MUST send multiple XOFF frames to
ensure that the DUT remains paused until the XON frame is
transmitted.

6.1.3. Procedure

The test tool SHOULD generate test traffic for at least 30 seconds
before sending any XOFF frame in order for the DUT to reach a steady-
state forwarding condition. The test tool then transmits one or more
XOFF frames on one or more ports. Each XOFF frame SHOULD instruct
the DUT to pause one or more of the Classification Groups currently
being forwarded by the DUT. The test tool MAY optionally send a XON
frame to instruct the DUT to resume transmission.

6.1.4. Measurements

The following measurements MUST be reported for each test port and
codepoint involved in the test.

Offered Load - the Offered Load from the DUT in N-octet frames per
second or bits per second. Note: The Offered Load from the DUT
may be insufficient to accurately measure the DUT's Pause Response
Time. This condition SHOULD be noted in the results.

The total number of PFC frames transmitted to the DUT by the test
tool.

The following values SHOULD be reported in either quanta OR
seconds:

Pause Response Time - The time between the transmit time of the
last bit of the pause frame and the receive time of the first
bit of the last codepoint matching test frame forwarded by the
DUT before the DUT is observed to pause the intended queue.

Intended Pause Time - The total time the test tool instructed
the DUT to pause.

Observed Pause Time - The actual time the DUT was observed to
pause.

XON Response Time - The time between the transmit time of the
last bit of the XON frame and the receive time of the first bit
of the first unpaused test packet from the DUT.

6.1.5. Reporting Format

   TBD

6.2. Queueput

6.2.1. Objective

   To determine the Queueput for one or more Traffic Classifications of
   a DUT using priority flow control.

6.2.2. Setup Parameters

   The following parameters MUST be defined. Each variable is
   configured with the following considerations.

      Each Classification Group MUST be listed. For each classification
      group, the following parameters MUST be specified:

         Codepoint - For DCB tests, the codepoint is the VLAN priority.

         Frame Size - The frame size includes both the CRC and VLAN
         header. See Section 5.1.4 for recommended frame sizes.

         Burst Size - The burst size specifies the number of frames
         transmitted with the minimum legal IFG before pausing. See
         Section 5.1.5.

         Intended Vector - The intended vector SHOULD specify the
         intended rate of test traffic specified as a percentage of port
         load.

         Traffic Pattern - The traffic distribution and traffic
         orientation used for this Classification.

         Conformance Vector - The conformance vector is optional, but
         MUST be defined if used.

      Priority-based Flow Control - PFC mechanisms MUST be enabled.

      Background Traffic - Background traffic MAY be present.

6.2.3. Procedure

   A search algorithm is used to determine the Queueput for each
   Classification Group. If Queue Congestion is detected for a
   Classification Group during a trial, then the Intended Vector for the
   Classification Group MUST be reduced for the subsequent trial. If a

Conformance Vector is specified for the test and Non-conformant
frames are received during a trial, then the Intended Vector SHOULD
be reduced for the subsequent trial. The algorithm MUST adjust the
Intended Vector for each Classification Group. The search algorithms
for each Classification Group MAY be run in parallel. The test
continues until all Classification Groups in the test have converged
on a discrete Queueput value.

6.2.4. Measurements

The Queueput for each Classification MUST be reported in either
N-octet frames per second or bits per second.

If a Conformance Vector is specified for a Classification Group, any
Non-conformant frames MUST be reported.

The number of PFC pause frames transmitted by the DUT for each code-
point in the Codepoint Set MUST be reported for each test port.

The total pause time observed by the tester for each code-point in
the Codepoint Set MUST be reported for each test port.

Any frame loss observed for test traffic using PFC enabled codepoints
MUST be reported. Any frame loss observed for test traffic using
non-PFC enabled codepoints on uncongested egress ports SHOULD be
reported, as that indicates the DUT is performing Head of Line
Blocking (HOLB).

6.2.5. Reporting Format

TBD

6.3. Maximum Forwarding Rate

6.3.1. Objective

To determine the maximum forwarding rate of one or more PFC queues on
a PFC capable DUT.

6.3.2. Setup Parameters

Maximum Forwarding Rate is conceptually similar to the measurement in
[RFC2285] but works on a per-Classification basis in a DCB context.
The following parameters MUST be defined. Each variable is
configured with the following considerations.

   Each Classification Group MUST be listed. For each classification
   group, the following parameters MUST be specified:

Codepoint - For DCB tests, the codepoint is the VLAN priority.

Frame Size - The frame size includes both the CRC and VLAN
header. See Section 5.1.4 for recommended frame sizes.

Burst Size - The burst size specifies the number of frames
transmitted with the minimum legal IFG before pausing. See
Section 5.1.5.

Intended Vector - The intended vector includes the intended
rate of test traffic specified as a percentage of port load.

Traffic Pattern - The traffic distribution and traffic
orientation used for this Classification.

Conformance Vector - The conformance vector is optional, but
MUST be defined if used.

Priority-based Flow Control - PFC mechanisms SHOULD be disabled.

Background Traffic - Background traffic MAY be present.

6.3.3. Procedure

The tester should iterate across all configured permutations of frame
size, burst size, and Intended Vector for all Classification Groups.

6.3.4. Measurements

The forwarding rate of each Classification Group MUST be reported as
the number of N-octet test frames per second the DUT correctly
forwards to the proper egress port.

The maximum forwarding rate for each Classification Group MUST be
reported as the highest recorded forwarding rate from the set of all
iterations.

Both the Intended and Offered Vector of each Classification Group
MUST be reported.

If a Conformance Vector is specified for a Classification Group, any
Non-conformant frames MUST be reported.

The number of PFC pause frames transmitted by the DUT for each code-
point in the Codepoint Set MUST be reported.

The total pause time observed by the tester for each code-point in
the Codepoint Set MUST be reported.

[6.3.5](#). **Reporting Format**

   TBD

[6.4](#). **Back-off**

[6.4.1](#). **Objective**

   To determine the delta between the maximum forwarding rate of a DUT
   and the point where the DUT ceases to use PFC to manage priority
   queues.

[6.4.2](#). **Setup Parameters**

   The following parameters MUST be defined. Each variable is
   configured with the following considerations.

      Each Classification Group MUST be listed. For each classification
      group, the following parameters MUST be specified:

         Codepoint - For DCB tests, the codepoint is the VLAN priority.

         Frame Size - The frame size includes both the CRC and VLAN
         header. See [Section 5.1.4](#) for recommended frame sizes.

         Burst Size - The burst size specifies the number of frames
         transmitted with the minimum legal IFG before pausing. See
         [Section 5.1.5](#).

         Intended Vector - The intended vector includes the intended
         rate of test traffic specified as a percentage of port load.

         Traffic Pattern - The traffic distribution and traffic
         orientation used for this Classification.

         Conformance Vector - The conformance vector is optional, but
         MUST be defined if used.

      Priority-based Flow Control - PFC mechanisms MUST be enabled.

      Backoff method - The recommended backoff method is to reduce the
      aggregate traffic load by a fixed amount while still maintaining a
      fixed load ratio between all Classification Groups.

6.4.3. Procedure

   The initial trial SHOULD begin with an Intended Load equal to or
   greater than the Maximum Forwarding Rate of the DUT/SUT. For each
   subsequent trial, the aggregate load is reduced until the DUT is
   observed to complete a trial without activating any Congestion
   Management methods.

6.4.4. Measurements

   The Intended and Offered Vector for each Classification Group MUST be
   reported.

   If a Conformance Vector is specified for a Classification Group, any
   Non-conformant frames MUST be reported.

   The number of PFC pause frames transmitted by the DUT for each code-
   point in the Codepoint Set MUST be reported.

   The total pause time observed by the tester for each code-point in
   the Codepoint Set MUST be reported.

   Any frame loss observed for test traffic using PFC enabled codepoints
   MUST be reported. Any frame loss observed for test traffic using
   non-PFC enabled codepoints on uncongested egress ports SHOULD be
   reported, as that indicates the DUT is performing Head of Line
   Blocking (HOLB).

6.4.5. Reporting Format

   TBD

6.5. Back-to-Back

6.5.1. Objective

   To determine the maximum duration a DUT can forward test traffic with
   minimum Inter-Frame Gap on one or more PFC queues without using
   Congestion Management.

6.5.2. Setup Parameters

   The following parameters MUST be defined. Each variable is
   configured with the following considerations

      Each Classification Group MUST be listed. For each classification
      group, the following parameters MUST be specified:

Codepoint - For DCB tests, the codepoint is the VLAN priority.

Frame Size - The frame size includes both the CRC and VLAN header. See Section 5.1.4 for recommended frame sizes.

Intended Vector - The intended vector includes the intended rate of test traffic specified as a percentage of port load.

Traffic Pattern - The traffic distribution and traffic orientation used for this Classification.

Conformance Vector - The conformance vector is optional, but MUST be defined if used.

Priority-based Flow Control - PFC mechanisms MUST be enabled.

The sum of all Intended Vectors on a transmitting port SHOULD equal the Maximum Offered Load (MOL) of that port.

6.5.3. Procedure

A search algorithm is used to determine the maximum duration in seconds for which the configured Classification Profile can be forwarded by the DUT without active Congestion Management. If Congestion Management is detected during an iteration, then the duration MUST be reduced for the next iteration.

6.5.4. Measurements

The Intended and Offered Vector for each Classification Group MUST be reported.

If a Conformance Vector is specified for a Classification Group, any Non-conformant frames MUST be reported.

The number of PFC pause frames transmitted by the DUT for each code-point in the Codepoint Set MUST be reported.

The total pause time observed by the tester for each code-point in the Codepoint Set MUST be reported.

Any frame loss observed for test traffic using PFC enabled codepoints MUST be reported. Any frame loss observed for test traffic using non-PFC enabled codepoints on uncongested egress ports SHOULD be reported, as that indicates the DUT is performing Head of Line Blocking (HOLB).

6.5.5. **Reporting Format**

   TBD

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 constraints
   specified in the sections above.

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

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

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

8. **IANA Consdierations**

   Testers SHOULD use network addresses assigned by IANA for the purpose
   of testing networks.

9. **Normative References**

   [RFC1242] Bradner, S., "Benchmarking terminology for network
              interconnection devices", RFC 1242, July 1991.

   [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate
              Requirement Levels", BCP 14, RFC 2119, March 1997.

   [RFC2285] Mandeville, R., "Benchmarking Terminology for LAN
              Switching Devices", RFC 2285, February 1998.

   [RFC2544] Bradner, S. and J. McQuaid, "Benchmarking Methodology for
              Network Interconnect Devices", RFC 2544, March 1999.

   [RFC2889] Mandeville, R. and J. Perser, "Benchmarking Methodology
              for LAN Switching Devices", RFC 2889, August 2000.

   [RFC3643]   Weber, R., Rajagopal, M., Travostino, F., O'Donnell, M.,
               Monia, C., and M. Merhar, "Fibre Channel (FC) Frame
               Encapsulation", RFC 3643, December 2003.

   [RFC4689]   Poretsky, S., Perser, J., Erramilli, S., and S. Khurana,
               "Terminology for Benchmarking Network-layer Traffic
               Control Mechanisms", RFC 4689, October 2006.

   [RFC5180]   Popoviciu, C., Hamza, A., Van de Velde, G., and D.
               Dugatkin, "IPv6 Benchmarking Methodology for Network
               Interconnect Devices", RFC 5180, May 2008.

Appendix A. Acknowledgements

Authors' Addresses

   Timmons C. Player
   Spirent Communications

   Email: timmons.player@spirent.com

   David Newman
   Network Test

   Email: dnewman@networktest.com