IPv6 Operations F. Baker
Internet-Draft Cisco Systems
Intended status: Informational November 7, 2010
Expires: May 11, 2011

Testing Eyeball Happiness
draft-baker-bmwg-testing-eyeball-happiness-00

Abstract

A barrier to the deployment of IPv6 is the amount of time it takes to open a session using common transport APIs in dual stack networks and networks with filtering such as proposed in BCP 38. This note describes a test that can be used by a manufacturer or network operator to determine whether an application adequately meets the "happy eyeballs" requirements. This test is not a test of a specific algorithm, but of the external behavior of the system as a black box. Any algorithm that has the intended external behavior will be accepted by it.

Status of this Memo

Copyright Notice

Table of Contents

1. Introduction

The Happy Eyeballs [I-D.wing-v6ops-happy-eyeballs-ipv6] draft observes on an issue in deployed IPv6-only and dual stack networks, and proposes a correction. [RFC5461] similarly looks at TCP's response to so-called "soft errors" (ICMP host and network unreachable messages), pointing out an issue and a set of solutions. In short, in a network that contains both IPv4 [RFC0791] and IPv6 [RFC2460] prefixes and routes, the fact that two hosts that need to communicate have an addresses using the same architecture doesn't imply that the network has usable routes connecting them, or that those addresses are useful to the applications in question. In addition, the process of opening a session using the Sockets API [RFC3493] is generally described in terms of obtaining a list of possible addresses for a peer (which will normally include both IPv4 and IPv6 addresses) using getaddrinfo() and trying them in sequence until one succeeds or all have failed. This naive algorithm, if implemented as described, has the side-effect of making the worst case delay in opening a session far longer than human patience normally allows.

This note describes a test that can be used by a manufacturer or network operator to determine whether an application adequately meets the "happy eyeballs" requirements. This test is not a test of a specific algorithm, but of the external behavior of the system as a black box. Any algorithm that has the intended external behavior will be accepted by it.

2. Generic Test

The proposed test assumes that the application works in an IPv4 network; the IPv4 option has to be part of the test. That question devolves to whether the application can open a session in a timely fashion. The issue that ISPs are reporting is that a host (MacOSX, Windows, Linux, FreeBSD, etc) that has more than one address (an IPv4 and an IPv6 address, two global IPv6 addresses, etc) may serially try addresses, allowing the TCP setup to expire (3 seconds or thereabouts) for each attempt. There have been reports of a session setup taking as long as 40 seconds as a result. In addition, at least Apple and apparently some versions of Windows wonder about A and AAAA records, and if there is a AAAA record try to use the indicated IPv6 address and *never*fail*over*to*IPv4*. As a result, there is at least one ISP that has told me that it can't advertise AAAA records for its mail services because the neighboring (and dominant) ISP runs IPv6 as a walled garden.

What I have proposed as a test is essentially this: consider two computers, Alice and Bob, as shown in Figure 1.

```
                  |192.0.2.0/24
          +-----+ |2001:DB8:1:0::/64     | +-----+
          |Alice+-+2001:DB8:0:1::/64     +-+ Bob |
          +-----+ | +-------+  +-------+ | +-----+
                  +-+Router1|  |Router2+-+
          +-----+ | +-----+-+  +-+-----+ |198.51.100.0/24
          | DNS +-+       |      |       |2001:DB8:0:2::/64
          +-----+ |      -+------+-      |2001:DB8:1:4::/64
                       203.0.113.0/24    |2001:DB8:2:4::/64
                       2001:DB8:0:3::/64
```

Figure 1: Generic Test Environment

Alice and Bob each have a set of one or more IPv4 and two or more IPv6 addresses in DNS, and the router is configured to route the relevant prefixes. The test plays with an ACL in the router that would prevent traffic Alice->Bob using each of Bob’s addresses. If Bob has a total of N addresses, we run the test N times, permitting exactly one of the addresses each time. The test is presumed to have passed if, on each attempt, the session can be set up within a stated interval, on the order of 500 ms perhaps.

Multiple routers are used to facilitate the use of null routing or the removal of routes in Router1 that Router would serve as local and therefore non-removable routes. In some operating systems, this can be simulated within a single router.

In addition, for some applications, a more elaborate test environment may be necessary. For example, when testing an application that uses IP multicast, it may be appropriate to provide multiple instances of Bob, perhaps on different LANs, in order to test the application behavior adequately. This is considered beyond the scope of this present note, as it is very specific to the application, but test engineers should ask themselves that question when designing a test for a new application.

2.1. In more detail

As initial conditions, as shown in Figure 1,

- o Alice, DNS, and Router1 each have addresses in 192.0.2.0/24, 2001:DB8:1:0::/64, and 2001:DB8:0:1::/64 on the same LAN,
- o Router1 and Router2 each have addresses in 203.0.113.0/24 and 2001:DB8:0:3::/64 on the same LAN,

- o Router2 and Bob have addresses in 198.51.100.0/24, 2001:DB8:0:2::/64, 2001:DB8:1:4::/64, and 2001:DB8:2:4::/64 on the same LAN,

- o Router1 has routes to 198.51.100.0/24 2001:DB8:0:2::/64 2001:DB8:1:4::/64 2001:DB8:2:4::/64 via Router2

- o Router2 has routes to 192.0.2.0/24, 2001:DB8:1:0::/64, and 2001:DB8:0:1::/64 via Router1,

- o DNS has appropriate A and AAAA records for Alice and Bob listing their addresses.

The means of accomplishing this configuration - static configuration of addresses and prefixes, DHCP/DHCPv6, and SLAAC, and the routing protocol or static route configuration - are irrelevant beyond noting them in the test report. If only DHCPv6 is tested, the test report should say so, for example.

In addition, there are three means of disrupting routes:

- o An ACL filter, configured to respond with no ICMP response

- o An ACL filter, configured to result in an ICMP "administratively unreachable"

- o A null or lacking route, which would result in an ICMP destination unreachable.

Alice is the unit under test. Most of the applications in real world obtain the addresses their correspondents from DNS. Therefore, the IPv4 and IPv6 addresses for Alice and Bob need to be stored within a test DNS server, and the communication done by name. The test is conducted several times with varying routing and filtering combinations, testing if not every combination of addresses, every combination of relevant condition sets. If the ordering received from DNS is deterministic, the test simply requires testing of each scenario. However, the order of the addresses within the DNS reply is not always deterministic; in such a case, there SHOULD be enough iterations of the test performed to ensure that the set of scenarios is adequately tested.

The test is first conducted with no disruptions. One should be able to observe the application working as expected between Alice and Bob; if it is a web service, for example, one should be able to load a web page, and if it is instant messaging, one should be able to have a breif conversation. Which set of addresses is chosen is irrelevant. What is important is an observation that the application works as expected under some set of sircumstances, and the duration from

Alice’s initial DNS request for Bob’s addresses to the arrival of Bob’s first application response at Alice.

Subsequent instances of the test should test a variety of scenarios including:

- o Bob is unreachable from Alice, for each of the various reasons, using IPv4.
- o Bob is unreachable from Alice, for each of the various reasons, using IPv6 at any address.
- o Bob is reachable from Alice using only one IPv6 address (testing each address assigned to Bob in the setup), with various kinds of blockage.

It would be worthwhile to go through the test once clearing all state in the UUT (Alice) between tests, and again ensuring that Alice is unaware of any changes in the network so that memory effects between tests can be explored. In at least one case, the DNS resource records obtained by Alice’s resolver should be permitted to time out, testing whether Alice will re-retreive them. The fact that Alice was able to contain Bob at a given address should not preclude Alice trying other addresses on subsequent attempts.

One would expect, in the worst case in an environment with nominal end to end delay, for an application to be set up in the time measured in the first instance of the test plus at most one inter-attempt interval per address that Bob has. One might allow an additional 50 ms for natural host variability. The [I-D.wing-v6ops-happy-eyeballs-ipv6] section 4.1, calls this "p*10" for some value of p, which must not exceed 4 seconds in the worst case. The test is considered to have been passed if, on each pass through the test, an application session succeeded in opening, and they each took approximately the same amount of time within the parameters of the Happy Eyeballs draft.

3. IANA Considerations

This memo asks the IANA for no new parameters.

Note to RFC Editor: This section will have served its purpose if it correctly tells IANA that no new assignments or registries are required, or if those assignments or registries are created during the RFC publication process. From the author"s perspective, it may therefore be removed upon publication as an RFC at the RFC Editor"s discretion.

4. Security Considerations

This note doesn't address security-related issues.

5. Acknowledgements

This note was discussed with Dan Wing, Andrew Yourtchenko, and Fernando Gont.

6. Change Log

-00 Version: Initial version - November, 2010

7. References

7.1. Normative References

[I-D.wing-v6ops-happy-eyeballs-ipv6]
Wing, D. and A. Yourtchenko, "Happy Eyeballs: Trending Towards Success with Dual-Stack Hosts", draft-wing-v6ops-happy-eyeballs-ipv6-01 (work in progress), October 2010.

[RFC5461] Gont, F., "TCP's Reaction to Soft Errors", RFC 5461, February 2009.

7.2. Informative References

[RFC0791] Postel, J., "Internet Protocol", STD 5, RFC 791, September 1981.

[RFC2460] Deering, S. and R. Hinden, "Internet Protocol, Version 6 (IPv6) Specification", RFC 2460, December 1998.

[RFC3493] Gilligan, R., Thomson, S., Bound, J., McCann, J., and W. Stevens, "Basic Socket Interface Extensions for IPv6", RFC 3493, February 2003.

Author's Address

Fred Baker
Cisco Systems
Santa Barbara, California 93117
USA

Email: fred@cisco.com

Internet Engineering Task Force M. Hamilton
Internet-Draft BreakingPoint Systems
Intended status: Informational S. Banks
Expires: April 28, 2011 Cisco Systems
October 25, 2010

Benchmarking Methodology for Content-Aware Network Devices
draft-hamilton-bmwg-ca-bench-meth-05

Abstract

The purpose of this document is to define a set of test scenarios which may be used to create a series of statistics that will help to better understand the performance of network devices that operate at nework layers above IP. More specifically, these scenarios are designed to most accurately predict performance of these devices when subjected to modern traffic patterns. This document will operate within the constraints of the Benchmarking Working Group charter, namely black box characterization in a laboratory environment.

Table of Contents

1. Introduction

Content-aware and deep packet inspection (DPI) device penetration has grown exponentially over the last decade. No longer are devices simply using Ethernet headers and IP headers to make forwarding decisions. Devices that could historically be classified as 'stateless' or raw forwarding devices are now seeing more DPI functionality. Devices such as core and edge routers are now being developed with DPI functionality to make more intelligent routing and forwarding decisions.

The Benchmarking Working Group (BMWG) has historically produced Internet Drafts and Requests for Comment that are focused specifically on creating output metrics that are derived from a very specific and well-defined set of input parameters that are completely and unequivocally reproducible from testbed to testbed. The end goal of such methodologies is to, in the words of the BMWG charter "reduce specmanship" from network equipment manufacturers(NEM's). Existing BMWG work has certainly met this stated goal.

Today, device sophistication has surpassed existing methodologies, allowing vendors to reengage in specmanship. In order to achieve the stated BMWG goals, the methodologies designed to hold vendors accountable must evolve with the enhanced device functionality.

The BMWG has historically avoided the use of the term "realistic" throughout all of its drafts and RFCs. While this document will not explicitly use this term, the spirit will remain. Admittedly, the term has an infinite number of definitions depending on the context or environment in which it is used.

The primary purpose of this document is not to replace existing methodologies, but to provide a more modern approach to benchmarking network devices that complements the data acquired using existing BMWG methodologies. Existing BMWG work generally revolves around completely repeatable input stimulus, expecting fully repeatable output. This document departs from this mantra, although utilizes some of the same principles. This methodology is more focused on output repeatability than on static input stimulus.

Many of the terms used throughout this draft have previously been defined in "Benchmarking Terminology for Firewall Performance" RFC 2647 [1]. This document SHOULD be consulted prior to using this document. The Benchmarking Methodology Working Group (BMWG) has previously defined methodologies for network interconnect devices with RFC 2544 [2] and firewall performance with RFC 3511 [3].

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 [4].

2. Scope

Content-aware devices take many forms, shapes and architectures. These devices are advanced network interconnect devices that inspect deep into the application payload of network data packets to do classification. They may be as simple as a firewall that uses application data inspection for rule set enforcement, or they may have advanced functionality such as performing protocol decoding and validation, anti-virus, anti-spam and even application exploit filtering.

It shall be explicitly stated that this methodology does not imply the use of traffic captured from live networks and replayed.

This document is strictly focused on examining performance and robustness across a focused set of metrics that may be used to more accurately predict device performance when deployed in modern networks. These metrics will be implementation independent.

It should also be noted that the purpose of this document is not to perform functional testing of the potential features in the Device/System Under Test (DUT/SUT)[1] nor specify the configurations that should be tested. Various definitions of proper operation and configuration may be appropriate within different contexts. While the definition of these parameters are outside the scope of this document, the specific configuration of both the DUT and tester SHOULD be published with the test results for repeatability and comparison purposes.

While a list of devices that fall under this category will quickly become obsolete, an initial list of devices that would be well served by utilizing this type of methodology should prove useful. Devices such as firewalls, intrusion detection and prevention devices, application delivery controllers, deep packet inspection devices, and unified threat management systems generally fall into the content-aware category.

3. Test Setup

This document will be applicable to most test configurations and will not be confined to a discussion on specific test configurations. Since each DUT/SUT will have their own unique configuration, users MUST configure their device with the same parameters that would be used in the actual deployment of the device. The DUT configuration MUST be published with the final benchmarking results. If available, command-line scripts used to configured the DUT SHOULD be published with the final results.

The lines between network boundaries are rapidly blurring. No longer are there just single and dual-homed devices; this methodology will be based on a fully meshed network topology. Organizations deploying content-aware devices are doing so throughout their network infrastructure. These devices inspect deep into the application flow to perform quality of service monitoring, filtering, metering, threat mitigation and more.

Figure 1 illustrates a network topology that is fully meshed.

```
                      +---+      +---+      +---+
                      |C/S|      |C/S|      |C/S|
                      +---+      +---+      +---+
                          \        |        /
                           \ +----------+ /
                            \|          |/
                    +---+____|   DUT/   |____+---+
                    |C/S|    |   SUT    |    |C/S|
                    +---+   /|          |\   +---+
                           / +----------+ \
                          /        |       \
                      +---+      +---+     +---+
                      |C/S|      |C/S|     |C/S|
                      +---+      +---+     +---+
```

Fully Meshed Device

Figure 1: Fully Meshed Device

3.1. Test Considerations

3.2. Clients and Servers

Content-aware device testing SHOULD involve multiple clients and multiple servers. As with RFC 3511 [3], this methodology will use the terms virtual clients/servers throughout. Similarly defined in RFC 3511 [3], a data source may emulate multiple clients and/or servers within the context of the same test scenario. The test report MUST indicate the number of virtual clients/servers used during the test. In Appendix C of RFC 2544 [2], the range of IP addresses assigned to the BMWG by the IANA are listed. This address range SHOULD be adhered to in accordance with RFC 2544 [2]. Additionally, section 5.2 of RFC 5180 [5] SHOULD be consulted for the appropriate address ranges when testing IPv6-enabled configurations.

3.3. Traffic Generation Requirements

The explicit purposes of content-aware devices vary widely, but these devices use information deeper inside the application flow to make decisions and classify traffic. This methodology will not utilize traffic flows representing application traffic, but will use the shells of these application flows for benchmarking purposes. The term "Application Flow" is defined in RFC 2722 [6]. Using the shell simply means sending arbitrary payload over the established session rather than actual application payload.

The test tool MUST be able to open TCP connections on multiple destination ports and MUST be able to direct UDP traffic to multiple destination ports. The transport layer payload SHOULD be alternating zeros and ones, but MAY be random.

This document will illustrate an example mix of what traffic may look like on a sample modern network, though the authors understand that no two networks look alike. If a user of this methodology understands the traffic patterns in their modern network, that user MAY use the framework for traffic specification to evaluate their DUT.

3.4. Framework for Traffic Specification

The following table MUST be specified for each application. In cases where there are multiple destination ports, they should be evenly distributed across.

- o Percentage of Total Bandwidth: 25%

- o Client Originated Flow Bandwidth: 15%

- o Server Originated Flow Bandwidth: 85%

- o Transport Protocol: TCP

- o Destination Port: 80

- o Average Layer 4 Flow Size: 256 kB

3.5. Multiple Client/Server Testing

In actual network deployments, connections are being established between multiple clients and multiple servers simultaneously. Device vendors have been known to optimize the operation of their devices for easily defined patterns. The connection sequence ordering scenarios a device will see on a network will likely be much less deterministic. Thus, users SHOULD setup the test equipment to issue requests at random to the virtual servers rather than in a predictable round-robin fashion. This method will help to appropriately reflect network deployment behavior in the test setup.

3.6. Network Address Translation

Many content-aware devices are capable of performing Network Address Translation (NAT)[1]. If the final deployment of the DUT will have this functionality enabled, then the DUT MUST also have it enabled during the execution of this methodology. It MAY be beneficial to perform the test series in both modes in order to determine the performance differential when using NAT. The test report MUST indicate whether NAT was enabled during the testing process.

3.7. TCP Stack Considerations

As with RFC 3511 [3], TCP options SHOULD remain constant across all devices under test in order to ensure truly comparable results. This document does not attempt to specify which TCP options should be used, but all devices tested SHOULD be subject to the same configuration options.

3.8. Other Considerations

Various content-aware devices will have widely varying feature sets. In the interest of representative test results, the DUT features that will likely be enabled in the final deployment SHOULD be used. This methodology is not intended to advise on which features should be enabled, but to suggest using actual deployment configurations.

4. Benchmarking Tests

4.1. Maximum Application Connection Establishment Rate

4.1.1. Objective

To determine the maximum rate through which a device is able to establish application-specific sessions as defined by RFC 2647 [1].

4.1.2. Setup Parameters

The following parameters MUST be defined for all tests:

4.1.2.1. Transport-Layer Parameters

- o Aging Time: The time, expressed in seconds that the DUT will keep a connection in its state table after receiving a TCP FIN or RST packet.

- o Maximum Segment Size: The size in bytes of the largest segment which may be sent over a TCP connection.

4.1.2.2. Application-Layer Parameters

For each application protocol in use during the test run, the table provided in Section 3.4 must be published.

4.1.3. Procedure

The test SHOULD generate application network traffic that meets the conditions of Section 3.3. The traffic pattern SHOULD begin with an application session establishment rate of 10% of expected maximum. The test SHOULD be configured to increase the attempt rate in units of 10 up through 110% of expected maximum. The duration of each loading phase SHOULD be at least 30 seconds. This test MAY be repeated, each subsequent iteration beginning at 5% of expected maximum and increasing session establishment rate to 10% more than the maximum observed from the previous test run.

This procedure MAY be repeated any number of times with the results being averaged together.

4.1.4. Measurement

The following metrics MAY be determined from this test, and SHOULD be observed for each application protocol within the traffic mix:

4.1.4.1. Maximum Application Connection Establishment Rate

The test tool SHOULD report the maximum rate at which application connections were established, as defined by RFC 2647 [1], Section 3.7. This rate SHOULD be reported individually for each application protocol present within the traffic mix.

4.1.4.2. Application Connection Setup Time

The test tool SHOULD report the minimum, maximum and average application setup time, as defined by RFC 2647 [1], Section 3.9. This rate SHOULD be reported individually for each application protocol present within the traffic mix.

4.1.4.3. Application Connection Response Time

The test tool SHOULD report the minimum, maximum and average application session response times. This metric is defined as the time between when the first SYN was sent and the arrival of the corresponding SYN-ACK. This metric does not apply for non connection-based protocols.

4.1.4.4. Application Connection Time To Close

The test tool SHOULD report the minimum, maximum and average application session time to close, as defined by RFC 2647 [1], Section 3.13. This rate SHOULD be reported individually for each application protocol present within the traffic mix.

4.1.4.5. Packet Loss

The test tool SHOULD report the number of network packets lost or dropped from source to destination.

4.1.4.6. Application Latency

The test tool SHOULD report the minimum, maximum and average amount of time an application packet takes to traverse the DUT, as defined by RFC 1242 [7], Section 3.13. This rate SHOULD be reported individually for each application protocol present within the traffic mix.

4.2. Application Throughput

4.2.1. Objective

To determine the maximum rate through which a device is able to forward bits when using stateful applications.

4.2.2. Setup Parameters

The following parameters MUST be defined and reported for all tests:

4.2.2.1. Parameters

The same transport and application parameters as described in Section 4.1.2 MUST be used.

4.2.3. Procedure

This test will attempt to send application data through the device at a session rate of 30% of the maximum established as observed in Section 4.1. This procedure MAY be repeated with the results from each iteration averaged together.

4.2.4. Measurement

The following metrics MAY be determined from this test, and SHOULD be observed for each application protocol within the traffic mix:

4.2.4.1. Maximum Throughput

The test tool SHOULD report the minimum, maximum and average application throughput.

4.2.4.2. Packet Loss

The test tool SHOULD report the number of network packets lost or dropped from source to destination.

4.2.4.3. Application Connection Setup Time

The test tool SHOULD report the minimum, maximum and average application setup time, as defined by RFC 2647 [1], Section 3.9. This rate SHOULD be reported individually for each application protocol present within the traffic mix.

4.2.4.4. Application Connection Response Time

The test tool SHOULD report the minimum, maximum and average application session response times. This metric is defined as the time between when the first SYN was sent and the arrival of the corresponding SYN-ACK. This metric does not apply for non-connection oriented protocols.

4.2.4.5. Application Connection Time To Close

The test tool SHOULD report the minimum, maximum and average application session time to close, as defined by RFC 2647 [1], Section 3.13. This rate SHOULD be reported individually for each application protocol present within the traffic mix.

4.2.4.6. Application Latency

The test tool SHOULD report the minimum, maximum and average amount of time an application packet takes to traverse the DUT, as defined by RFC 1242 [7], Section 3.13. This rate SHOULD be reported individually for each application protocol present within the traffic mix.

4.3. Malicious Traffic Handling

4.3.1. Objective

To determine the effects on performance that malicious traffic may have on the DUT. While this test is not designed to characterize accuracy of detection or classification, it MAY be useful to record these measurements as specified below.

4.3.2. Setup Parameters

The same parameters must be used for Transport-Layer and Application Layer Parameters previously specified in Section 4.1.2 and Section 4.2.2, respectively. Additionally, the following parameters MUST be defined and reported for all tests:

o Attack List: A listing of the malicious traffic that was generated by the test.

4.3.3. Procedure

This test will utilize the procedures specified previously in Section 4.1.3 and Section 4.2.3. When performing the procedures listed previously, during the steady-state time, the tester should generate malicious traffic representative of the final network deployment. The mix of attacks MAY include software vulnerability exploits, network worms, back-door access attempts, network probes and other malicious traffic.

If a DUT may be run with and without the attack mitigation, both procedures SHOULD be run with and without the feature enabled on the DUT to determine the affects of the malicious traffic on the baseline metrics previously derived. If a DUT does not have active attack mitigation capabilities, this procedure SHOULD be run regardless. Certain malicious traffic could affect device performance even if the DUT does not actively inspect packet data for malicious traffic.

4.3.4. Measurement

The metrics specified by Section 4.1.4 and Section 4.2.4 SHOULD be determined from this test.

4.4. Malformed Traffic Handling

4.4.1. Objective

To determine the effects on performance and stability that malformed traffic may have on the DUT.

4.4.2. Setup Parameters

The same parameters must be used for Transport-Layer and Application Layer Parameters previously specified in Section 4.1.2 and Section 4.2.2.

4.4.3. Procedure

This test will utilize the procedures specified previously in Section 4.1.3 and Section 4.2.3. When performing the procedures listed previously, during the steady-state time, the tester should generate malformed traffic at all protocol layers. This is commonly known as fuzzed traffic. Fuzzing techniques generally modify portions of packets, including checksum errors, invalid protocol options, and improper protocol conformance. This test SHOULD be run on a DUT regardless of whether it has built-in mitigation capabilities.

4.4.4. Measurement

For each protocol present in the traffic mix, the metrics specified by Section 4.1.4 and Section 4.2.4 MAY be determined. This data may be used to ascertain the effects of fuzzed traffic on the DUT.

5. Appendix A: Example Test Case

This appendix shows an example case of a protocol mix that may be used with this methodology.

Protocol	Label	Value
Web		
	Total BW	50%
	Client BW	15%
	Server BW	85%
	Transport Protocol	TCP
	Destination Port(s)	80
	Flow Size	256 kB
BitTorrent		
	Total BW	25%
	Client BW	2%
	Server BW	98%
	Transport Protocol	TCP
	Destination Port(s)	6881-6889
	Flow Size	150 MB
SMTP Email		
	Total BW	10%
	Client BW	90%
	Server BW	10%
	Transport Protocol	TCP
	Destination Port(s)	25
	Flow Size	40 kB
IMAP Email		
	Total BW	5%
	Client BW	20%
	Server BW	80%
	Transport Protocol	TCP
	Destination Port(s)	143
	Flow Size	30 kB
DNS		
	Total BW	5%
	Client BW	50%
	Server BW	50%
	Transport Protocol	UDP
	Destination Port(s)	53
	Flow Size	2 kB
RTP		
	Total BW	5%
	Client BW	1%
	Server BW	99%
	Transport Protocol	UDP
	Destination Port(s)	20000-65000
	Flow Size	100 MB

Table 1: Sample Traffic Pattern

6. IANA Considerations

This memo includes no request to IANA.

All drafts are required to have an IANA considerations section (see the update of RFC 2434 [8] for a guide). If the draft does not require IANA to do anything, the section contains an explicit statement that this is the case (as above). If there are no requirements for IANA, the section will be removed during conversion into an RFC by the RFC Editor.

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 RFC 2544 [2].

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

8. References

8.1. Normative References

[1] Newman, D., "Benchmarking Terminology for Firewall Performance", RFC 2647, August 1999.

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

[3] Hickman, B., Newman, D., Tadjudin, S., and T. Martin, "Benchmarking Methodology for Firewall Performance", RFC 3511, April 2003.

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

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

[6] Brownlee, N., Mills, C., and G. Ruth, "Traffic Flow Measurement: Architecture", RFC 2722, October 1999.

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

8.2. Informative References

[8] Narten, T. and H. Alvestrand, "Guidelines for Writing an IANA Considerations Section in RFCs", BCP 26, RFC 5226, May 2008.

Authors' Addresses

Mike Hamilton
BreakingPoint Systems
Austin, TX 78717
US

Phone: +1 512 636 2303
Email: mhamilton@breakingpoint.com

Sarah Banks
Cisco Systems
San Jose, CA 95134
US

Email: sabanks@cisco.com

Network Working Group S. Poretsky
Internet-Draft Allot Communications
Intended status: Informational B. Imhoff
Expires: May 12, 2011 Juniper Networks
K. Michielsen
Cisco Systems
November 8, 2010

Benchmarking Methodology for Link-State IGP Data Plane Route Convergence
draft-ietf-bmwg-igp-dataplane-conv-meth-22

Abstract

This document describes the methodology for benchmarking Link-State Interior Gateway Protocol (IGP) Route Convergence. The methodology is to be used for benchmarking IGP convergence time through externally observable (black box) data plane measurements. The methodology can be applied to any link-state IGP, such as ISIS and OSPF.

Status of this Memo

Copyright Notice

Table of Contents

1. Introduction and Scope

This document describes the methodology for benchmarking Link-State Interior Gateway Protocol (IGP) convergence. The motivation and applicability for this benchmarking is described in [Po09a]. The terminology to be used for this benchmarking is described in [Po10t].

IGP convergence time is measured on the data plane at the Tester by observing packet loss through the DUT. All factors contributing to convergence time are accounted for by measuring on the data plane, as discussed in [Po09a]. The test cases in this document are black-box tests that emulate the network events that cause convergence, as described in [Po09a].

The methodology described in this document can be applied to IPv4 and IPv6 traffic and link-state IGPs such as ISIS [Ca90][Ho08], OSPF [Mo98][Co08], and others.

2. Existing Definitions

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 BCP 14, RFC 2119 [Br97]. RFC 2119 defines the use of these key words to help make the intent of standards track documents as clear as possible. While this document uses these keywords, this document is not a standards track document.

This document uses much of the terminology defined in [Po10t] and uses existing terminology defined in other BMWG work. Examples include, but are not limited to:

```
Throughput                          [Ref.[Br91], section 3.17]
Device Under Test (DUT)             [Ref.[Ma98], section 3.1.1]
System Under Test (SUT)             [Ref.[Ma98], section 3.1.2]
Out-of-Order Packet                 [Ref.[Po06], section 3.3.4]
Duplicate Packet                    [Ref.[Po06], section 3.3.5]
Stream                              [Ref.[Po06], section 3.3.2]
Loss Period                         [Ref.[Ko02], section 4]
Forwarding Delay                    [Ref.[Po06], section 3.2.4]
IP Packet Delay Variation (IPDV)    [Ref.[De02], section 1.2]
```

3. Test Topologies

3.1. Test topology for local changes

Figure 1 shows the test topology to measure IGP convergence time due to local Convergence Events such as Local Interface failure (Section 8.1.1), layer 2 session failure (Section 8.2.1), and IGP adjacency failure (Section 8.2.2). This topology is also used to measure IGP convergence time due to the route withdrawal (Section 8.2.3), and route cost change (Section 8.3.2) Convergence Events. IGP adjacencies MUST be established between Tester and DUT, one on the Preferred Egress Interface and one on the Next-Best Egress Interface. For this purpose the Tester emulates two routers, each establishing one adjacency with the DUT. An IGP adjacency SHOULD be established on the Ingress Interface between Tester and DUT.

```
            ---------       Ingress Interface         ----------
            |       |<--------------------------------|        |
            |       |                                 |        |
            |       |    Preferred Egress Interface   |        |
            |  DUT  |-------------------------------->| Tester |
            |       |                                 |        |
            |       |-------------------------------->|        |
            |       |    Next-Best Egress Interface   |        |
            ---------                                 ----------
```

Figure 1: IGP convergence test topology for local changes

Figure 2 shows the test topology to measure IGP convergence time due to local Convergence Events with a non-ECMP Preferred Egress Interface and ECMP Next-Best Egress Interfaces (Section 8.1.1). In this topology, the DUT is configured with each Next-Best Egress interface as a member of a single ECMP set. The Preferred Egress Interface is not a member of an ECMP set. The Tester emulates N+1 next-hop routers, one router for the Preferred Egress Interface and N routers for the members of the ECMP set. IGP adjacencies MUST be established between Tester and DUT, one on the Preferred Egress Interface, an one on each member of the ECMP set. For this purpose each of the N+1 routers emulated by the Tester establishes one adjacency with the DUT. An IGP adjacency SHOULD be established on the Ingress Interface between Tester and DUT. When the test specifies to observe the Next-Best Egress Interface statistics, the combined statistics for all ECMP members should be observed.

```
            ---------       Ingress Interface         ----------
            |       |<--------------------------------|        |
            |       |    Preferred Egress Interface   |        |
            |       |-------------------------------->|        |
            |       |     ECMP set interface 1        |        |
            |  DUT  |-------------------------------->| Tester |
            |       |               .                 |        |
            |       |               .                 |        |
            |       |-------------------------------->|        |
            |       |     ECMP set interface N        |        |
            ---------                                 ----------
```

Figure 2: IGP convergence test topology for local changes with non-ECMP to ECMP convergence

3.2. Test topology for remote changes

Figure 3 shows the test topology to measure IGP convergence time due to Remote Interface failure (Section 8.1.2). In this topology the two routers R1 and R2 are considered System Under Test (SUT) and SHOULD be identically configured devices of the same model. IGP adjacencies MUST be established between Tester and SUT, one on the Preferred Egress Interface and one on the Next-Best Egress Interface. For this purpose the Tester emulates one or two routers. An IGP adjacency SHOULD be established on the Ingress Interface between Tester and SUT. In this topology there is a possibility of a transient microloop between R1 and R2 during convergence.

```
                       ------                      ----------
                       |    |  Preferred           |        |
              ------   | R2 |--------------------->|        |
              |    |-->|    |  Egress Interface    |        |
              |    |   ------                      |        |
              | R1 |                               | Tester |
              |    |           Next-Best           |        |
              |    |------------------------------>|        |
              ------           Egress Interface    |        |
                 ^                                 ----------
                 |                                     |
                 ---------------------------------------
                             Ingress Interface
```

Figure 3: IGP convergence test topology for remote changes

Figure 4 shows the test topology to measure IGP convergence time due to remote Convergence Events with a non-ECMP Preferred Egress Interface and ECMP Next-Best Egress Interfaces (Section 8.1.2). In this topology the two routers R1 and R2 are considered System Under

Test (SUT) and MUST be identically configured devices of the same model. Router R1 is configured with each Next-Best Egress interface as a member of the same ECMP set. The Preferred Egress Interface of R1 is not a member of an ECMP set. The Tester emulates N+1 next-hop routers, one for R2 and one for each member of the ECMP set. IGP adjacencies MUST be established between Tester and SUT, one on each egress interface of SUT. For this purpose each of the N+1 routers emulated by the Tester establishes one adjacency with the SUT. An IGP adjacency SHOULD be established on the Ingress Interface between Tester and SUT. In this topology there is a possibility of a transient microloop between R1 and R2 during convergence. When the test specifies to observe the Next-Best Egress Interface statistics, the combined statistics for all ECMP members should be observed.

```
                                        ------     ----------
                                        |    |     |        |
              ------      Preferred     | R2 |---->|        |
              |    |-------------------->|    |     |        |
              |    |  Egress Interface  ------     |        |
              |    |                               |        |
              |    |      ECMP set interface 1     |        |
              | R1 |------------------------------>| Tester |
              |    |               .               |        |
              |    |               .               |        |
              |    |               .               |        |
              |    |------------------------------>|        |
              ------      ECMP set interface N     |        |
                 ^                                 ----------
                 |                                     |
                 ---------------------------------------
                            Ingress Interface
```

Figure 4: IGP convergence test topology for remote changes with non-ECMP to ECMP convergence

3.3. Test topology for local ECMP changes

Figure 5 shows the test topology to measure IGP convergence time due to local Convergence Events of a member of an Equal Cost Multipath (ECMP) set (Section 8.1.3). In this topology, the DUT is configured with each egress interface as a member of a single ECMP set and the Tester emulates N next-hop routers, one router for each member. IGP adjacencies MUST be established between Tester and DUT, one on each member of the ECMP set. For this purpose each of the N routers emulated by the Tester establishes one adjacency with the DUT. An IGP adjacency SHOULD be established on the Ingress Interface between Tester and DUT. When the test specifies to observe the Next-Best Egress Interface statistics, the combined statistics for all ECMP members except the one affected by the Convergence Event, should be observed.

```
         ---------       Ingress Interface        ----------
         |       |<--------------------------------|        |
         |       |                                 |        |
         |       |     ECMP set interface 1        |        |
         |       |-------------------------------->|        |
         |  DUT  |                .                | Tester |
         |       |                .                |        |
         |       |                .                |        |
         |       |-------------------------------->|        |
         |       |     ECMP set interface N        |        |
         ---------                                  ----------
```

Figure 5: IGP convergence test topology for local ECMP changes

3.4. Test topology for remote ECMP changes

Figure 6 shows the test topology to measure IGP convergence time due to remote Convergence Events of a member of an Equal Cost Multipath (ECMP) set (Section 8.1.4). In this topology the two routers R1 and R2 are considered System Under Test (SUT) and MUST be identically configured devices of the same model. Router R1 is configured with each egress interface as a member of a single ECMP set and the Tester emulates N next-hop routers, one router for each member. IGP adjacencies MUST be established between Tester and SUT, one on each egress interface of SUT. For this purpose each of the N routers emulated by the Tester establishes one adjacency with the SUT. An IGP adjacency SHOULD be established on the Ingress Interface between Tester and SUT. In this topology there is a possibility of a transient microloop between R1 and R2 during convergence. When the test specifies to observe the Next-Best Egress Interface statistics, the combined statistics for all ECMP members except the one affected by the Convergence Event, should be observed.

```
                                        ------     ----------
                                        |    |     |        |
              ------      ECMP set      | R2 |---->|        |
              |    |------------------->|    |     |        |
              |    |      Interface 1   ------     |        |
              |    |                               |        |
              |    |      ECMP set interface 2     |        |
              | R1 |------------------------------>| Tester |
              |    |               .               |        |
              |    |               .               |        |
              |    |               .               |        |
              |    |------------------------------>|        |
              ------      ECMP set interface N     |        |
                 ^                                 ----------
                 |                                     |
                 ---------------------------------------
                            Ingress Interface
```

Figure 6: IGP convergence test topology for remote ECMP changes

3.5. Test topology for Parallel Link changes

Figure 7 shows the test topology to measure IGP convergence time due to local Convergence Events with members of a Parallel Link (Section 8.1.5). In this topology, the DUT is configured with each egress interface as a member of a Parallel Link and the Tester emulates the single next-hop router. IGP adjacencies MUST be established on all N members of the Parallel Link between Tester and DUT. For this purpose the router emulated by the Tester establishes N adjacencies with the DUT. An IGP adjacency SHOULD be established on the Ingress Interface between Tester and DUT. When the test specifies to observe the Next-Best Egress Interface statistics, the combined statistics for all Parallel Link members except the one affected by the Convergence Event, should be observed.

```
            ---------       Ingress Interface        ----------
            |       |<--------------------------------|        |
            |       |                                 |        |
            |       |     Parallel Link Interface 1   |        |
            |       |-------------------------------->|        |
            |  DUT  |              .                  | Tester |
            |       |              .                  |        |
            |       |              .                  |        |
            |       |-------------------------------->|        |
            |       |     Parallel Link Interface N   |        |
            ---------                                 ----------
```

Figure 7: IGP convergence test topology for Parallel Link changes

4. Convergence Time and Loss of Connectivity Period

Two concepts will be highlighted in this section: convergence time and loss of connectivity period.

The Route Convergence [Po10t] time indicates the period in time between the Convergence Event Instant [Po10t] and the instant in time the DUT is ready to forward traffic for a specific route on its Next-Best Egress Interface and maintains this state for the duration of the Sustained Convergence Validation Time [Po10t]. To measure Route Convergence time, the Convergence Event Instant and the traffic received from the Next-Best Egress Interface need to be observed.

The Route Loss of Connectivity Period [Po10t] indicates the time during which traffic to a specific route is lost following a Convergence Event until Full Convergence [Po10t] completes. This Route Loss of Connectivity Period can consist of one or more Loss Periods [Ko02]. For the testcases described in this document it is expected to have a single Loss Period. To measure Route Loss of Connectivity Period, the traffic received from the Preferred Egress Interface and the traffic received from the Next-Best Egress Interface need to be observed.

The Route Loss of Connectivity Period is most important since that has a direct impact on the network user’s application performance.

In general the Route Convergence time is larger than or equal to the Route Loss of Connectivity Period. Depending on which Convergence Event occurs and how this Convergence Event is applied, traffic for a route may still be forwarded over the Preferred Egress Interface after the Convergence Event Instant, before converging to the Next-Best Egress Interface. In that case the Route Loss of Connectivity Period is shorter than the Route Convergence time.

At least one condition needs to be fulfilled for Route Convergence time to be equal to Route Loss of Connectivity Period. The condition is that the Convergence Event causes an instantaneous traffic loss for the measured route. A fiber cut on the Preferred Egress Interface is an example of such a Convergence Event.

A second condition applies to Route Convergence time measurements based on Connectivity Packet Loss [Po10t]. This second condition is that there is only a single Loss Period during Route Convergence. For the testcases described in this document this is expected to be the case.

4.1. Convergence Events without instant traffic loss

To measure convergence time benchmarks for Convergence Events caused by a Tester, such as an IGP cost change, the Tester MAY start to discard all traffic received from the Preferred Egress Interface at the Convergence Event Instant, or MAY separately observe packets received from the Preferred Egress Interface prior to the Convergence Event Instant. This way these Convergence Events can be treated the same as Convergence Events that cause instantaneous traffic loss.

To measure convergence time benchmarks without instantaneous traffic loss (either real or induced by the Tester) at the Convergence Event Instant, such as a reversion of a link failure Convergence Event, the Tester SHALL only observe packet statistics on the Next-Best Egress Interface. If using the Rate-Derived method to benchmark convergence times for such Convergence Events, the Tester MUST collect a timestamp at the Convergence Event Instant. If using a loss-derived method to benchmark convergence times for such Convergence Events, the Tester MUST measure the period in time between the Start Traffic Instant and the Convergence Event Instant. To measure this period in time the Tester can collect timestamps at the Start Traffic Instant and the Convergence Event Instant.

The Convergence Event Instant together with the receive rate observations on the Next-Best Egress Interface allow to derive the convergence time benchmarks using the Rate-Derived Method [Po10t].

By observing lost packets on the Next-Best Egress Interface only, the observed packet loss is the number of lost packets between Traffic Start Instant and Convergence Recovery Instant. To measure convergence times using a loss-derived method, packet loss between the Convergence Event Instant and the Convergence Recovery Instant is needed. The time between Traffic Start Instant and Convergence Event Instant must be accounted for. An example may clarify this.

Figure 8 illustrates a Convergence Event without instantaneous traffic loss for all routes. The top graph shows the Forwarding Rate over all routes, the bottom graph shows the Forwarding Rate for a single route Rta. Some time after the Convergence Event Instant, Forwarding Rate observed on the Preferred Egress Interface starts to decrease. In the example, route Rta is the first route to experience packet loss at time Ta. Some time later, the Forwarding Rate observed on the Next-Best Egress Interface starts to increase. In the example, route Rta is the first route to complete convergence at time Ta’.

```
                ^
           Fwd  |
           Rate |-------------                    ............
                |             \                  .
                |              \                .
                |               \              .
                |                \            .
                |.................-.-.-.-.-.-.----------------
                +----+-------+---------------+----------------->
                ^    ^       ^               ^             time
               T0   CEI      Ta              Ta'

                ^
           Fwd  |
           Rate |-------------                .................
           Rta  |            |               .
                |            |               .
                |.............-.-.-.-.-.-.-.-.----------------
                +----+-------+---------------+----------------->
                ^    ^       ^               ^             time
               T0   CEI      Ta              Ta'

                Preferred Egress Interface: ---
                Next-Best Egress Interface: ...
```

With T0 the Start Traffic Instant; CEI the Convergence Event Instant; Ta the time instant traffic loss for route Rta starts; Ta' the time instant traffic loss for route Rta ends.

Figure 8

If only packets received on the Next-Best Egress Interface are observed, the duration of the packet loss period for route Rta can be calculated from the received packets as in Equation 1. Since the Convergence Event Instant is the start time for convergence time measurement, the period in time between T0 and CEI needs to be subtracted from the calculated result to become the convergence time, as in Equation 2.

```
Next-Best Egress Interface packet loss period
    = (packets transmitted
        - packets received from Next-Best Egress Interface) / tx rate
    = Ta' - T0
```

Equation 1

```
      convergence time
          = Next-Best Egress Interface packet loss period - (CEI - T0)
          = Ta' - CEI
```

Equation 2

4.2. Loss of Connectivity

Route Loss of Connectivity Period SHOULD be measured using the Route-Specific Loss-Derived Method. Since the start instant and end instant of the Route Loss of Connectivity Period can be different for each route, these can not be accurately derived by only observing global statistics over all routes. An example may clarify this.

Following a Convergence Event, route Rta is the first route for which packet loss starts, the Route Loss of Connectivity Period for route Rta starts at time Ta. Route Rtb is the last route for which packet loss starts, the Route Loss of Connectivity Period for route Rtb starts at time Tb with Tb>Ta.

```
                   ^
            Fwd    |
            Rate   |--------                       -----------
                   |        \                     /
                   |         \                   /
                   |          \                 /
                   |           \               /
                   |            ---------------
                   +------------------------------------------>
                            ^   ^             ^     ^      time
                           Ta   Tb           Ta'   Tb'
                                             Tb''  Ta''
```

Figure 9: Example Route Loss Of Connectivity Period

If the DUT implementation would be such that Route Rta would be the first route for which traffic loss ends at time Ta' with Ta'>Tb. Route Rtb would be the last route for which traffic loss ends at time Tb' with Tb'>Ta'. By using only observing global traffic statistics over all routes, the minimum Route Loss of Connectivity Period would be measured as Ta'-Ta. The maximum calculated Route Loss of Connectivity Period would be Tb'-Ta. The real minimum and maximum Route Loss of Connectivity Periods are Ta'-Ta and Tb'-Tb. Illustrating this with the numbers Ta=0, Tb=1, Ta'=3, and Tb'=5, would give a LoC Period between 3 and 5 derived from the global traffic statistics, versus the real LoC Period between 3 and 4.

If the DUT implementation would be such that route Rtb would be the first for which packet loss ends at time Tb'' and route Rta would be the last for which packet loss ends at time Ta'', then the minimum and maximum Route Loss of Connectivity Periods derived by observing only global traffic statistics would be Tb''-Ta, and Ta''-Ta. The real minimum and maximum Route Loss of Connectivity Periods are Tb''-Tb and Ta''-Ta. Illustrating this with the numbers Ta=0, Tb=1, Ta''=5, Tb''=3, would give a LoC Period between 3 and 5 derived from the global traffic statistics, versus the real LoC Period between 2 and 5.

The two implementation variations in the above example would result in the same derived minimum and maximum Route Loss of Connectivity Periods when only observing the global packet statistics, while the real Route Loss of Connectivity Periods are different.

5. Test Considerations

5.1. IGP Selection

The test cases described in Section 8 MAY be used for link-state IGPs, such as ISIS or OSPF. The IGP convergence time test methodology is identical.

5.2. Routing Protocol Configuration

The obtained results for IGP convergence time may vary if other routing protocols are enabled and routes learned via those protocols are installed. IGP convergence times SHOULD be benchmarked without routes installed from other protocols.

5.3. IGP Topology

The Tester emulates a single IGP topology. The DUT establishes IGP adjacencies with one or more of the emulated routers in this single IGP topology emulated by the Tester. See test topology details in Section 3. The emulated topology SHOULD only be advertised on the DUT egress interfaces.

The number of IGP routes and number of nodes in the topology, and the type of topology will impact the measured IGP convergence time. To obtain results similar to those that would be observed in an operational network, it is RECOMMENDED that the number of installed routes and nodes closely approximate that of the network (e.g. thousands of routes with tens or hundreds of nodes).

The number of areas (for OSPF) and levels (for ISIS) can impact the benchmark results.

5.4. Timers

There are timers that may impact the measured IGP convergence times. The benchmark metrics MAY be measured at any fixed values for these timers. To obtain results similar to those that would be observed in an operational network, it is RECOMMENDED to configure the timers with the values as configured in the operational network.

Examples of timers that may impact measured IGP convergence time include, but are not limited to:

Interface failure indication

IGP hello timer

IGP dead-interval or hold-timer

LSA or LSP generation delay

LSA or LSP flood packet pacing

SPF delay

5.5. Interface Types

All test cases in this methodology document MAY be executed with any interface type. The type of media may dictate which test cases may be executed. Each interface type has a unique mechanism for detecting link failures and the speed at which that mechanism operates will influence the measurement results. All interfaces MUST be the same media and Throughput [Br91][Br99] for each test case. All interfaces SHOULD be configured as point-to-point.

5.6. Offered Load

The Throughput of the device, as defined in [Br91] and benchmarked in [Br99] at a fixed packet size, needs to be determined over the preferred path and over the next-best path. The Offered Load SHOULD be the minimum of the measured Throughput of the device over the primary path and over the backup path. The packet size is selectable and MUST be recorded. Packet size is measured in bytes and includes the IP header and payload.

The destination addresses for the Offered Load MUST be distributed such that all routes or a statistically representative subset of all routes are matched and each of these routes is offered an equal share of the Offered Load. It is RECOMMENDED to send traffic matching all routes, but a statistically representative subset of all routes can be used if required.

In the Remote Interface failure testcases using topologies 3, 4, and 6 there is a possibility of a transient microloop between R1 and R2 during convergence. The TTL or Hop Limit value of the packets sent by the Tester may influence the benchmark measurements since it determines which device in the topology may send an ICMP Time Exceeded Message for looped packets.

The duration of the Offered Load MUST be greater than the convergence time plus the Sustained Convergence Validation Time.

Offered load should send a packet to each destination before sending another packet to the same destination. It is RECOMMENDED that the packets are transmitted in a round-robin fashion with a uniform interpacket delay.

5.7. Measurement Accuracy

Since packet loss is observed to measure the Route Convergence Time, the time between two successive packets offered to each individual route is the highest possible accuracy of any packet loss based measurement. The higher the traffic rate offered to each route the higher the possible measurement accuracy.

Also see Section 6 for method-specific measurement accuracy.

5.8. Measurement Statistics

The benchmark measurements may vary for each trial, due to the statistical nature of timer expirations, cpu scheduling, etc. Evaluation of the test data must be done with an understanding of generally accepted testing practices regarding repeatability, variance and statistical significance of a small number of trials.

5.9. Tester Capabilities

It is RECOMMENDED that the Tester used to execute each test case has the following capabilities:

1. Ability to establish IGP adjacencies and advertise a single IGP topology to one or more peers.

2. Ability to measure Forwarding Delay, Duplicate Packets and Out-of-Order Packets.

3. An internal time clock to control timestamping, time measurements, and time calculations.

4. Ability to distinguish traffic load received on the Preferred and Next-Best Interfaces [Po10t].

5. Ability to disable or tune specific Layer-2 and Layer-3 protocol functions on any interface(s).

The Tester MAY be capable to make non-data plane convergence observations and use those observations for measurements. The Tester MAY be capable to send and receive multiple traffic Streams [Po06].

Also see Section 6 for method-specific capabilities.

6. Selection of Convergence Time Benchmark Metrics and Methods

Different convergence time benchmark methods MAY be used to measure convergence time benchmark metrics. The Tester capabilities are important criteria to select a specific convergence time benchmark method. The criteria to select a specific benchmark method include, but are not limited to:

Tester capabilities:	Sampling Interval, number of Stream statistics to collect
Measurement accuracy:	Sampling Interval, Offered Load, number of routes
Test specification:	number of routes
DUT capabilities:	Throughput, IP Packet Delay Variation

6.1. Loss-Derived Method

6.1.1. Tester capabilities

The Offered Load SHOULD consist of a single Stream [Po06]. If sending multiple Streams, the measured packet loss statistics for all Streams MUST be added together.

In order to verify Full Convergence completion and the Sustained Convergence Validation Time, the Tester MUST measure Forwarding Rate each Packet Sampling Interval.

The total number of packets lost between the start of the traffic and the end of the Sustained Convergence Validation Time is used to calculate the Loss-Derived Convergence Time.

6.1.2. Benchmark Metrics

The Loss-Derived Method can be used to measure the Loss-Derived Convergence Time, which is the average convergence time over all routes, and to measure the Loss-Derived Loss of Connectivity Period, which is the average Route Loss of Connectivity Period over all routes.

6.1.3. Measurement Accuracy

The actual value falls within the accuracy interval [-(number of destinations/Offered Load), +(number of destinations/Offered Load)] around the value as measured using the Loss-Derived Method.

6.2. Rate-Derived Method

6.2.1. Tester Capabilities

The Offered Load SHOULD consist of a single Stream. If sending multiple Streams, the measured traffic rate statistics for all Streams MUST be added together.

The Tester measures Forwarding Rate each Sampling Interval. The Packet Sampling Interval influences the observation of the different convergence time instants. If the Packet Sampling Interval is large compared to the time between the convergence time instants, then the different time instants may not be easily identifiable from the Forwarding Rate observation. The presence of IPDV [De02] may cause fluctuations of the Forwarding Rate observation and can prevent correct observation of the different convergence time instants.

The Packet Sampling Interval MUST be larger than or equal to the time between two consecutive packets to the same destination. For maximum accuracy the value for the Packet Sampling Interval SHOULD be as small as possible, but the presence of IPDV may enforce using a larger Packet Sampling Interval. The Packet Sampling Interval MUST be reported.

IPDV causes fluctuations in the number of received packets during each Packet Sampling Interval. To account for the presence of IPDV in determining if a convergence instant has been reached, Forwarding Delay SHOULD be observed during each Packet Sampling Interval. The minimum and maximum number of packets expected in a Packet Sampling Interval in presence of IPDV can be calculated with Equation 3.

```
number of packets expected in a Packet Sampling Interval
  in presence of IP Packet Delay Variation
    = expected number of packets without IP Packet Delay Variation
      +/-( (maxDelay - minDelay) * Offered Load)
with minDelay and maxDelay the minimum resp. maximum Forwarding Delay
  of packets received during the Packet Sampling Interval
```

Equation 3

To determine if a convergence instant has been reached the number of packets received in a Packet Sampling Interval is compared with the range of expected number of packets calculated in Equation 3.

6.2.2. Benchmark Metrics

The Rate-Derived Method SHOULD be used to measure First Route Convergence Time and Full Convergence Time. It SHOULD NOT be used to measure Loss of Connectivity Period (see Section 4).

6.2.3. Measurement Accuracy

The measurement accuracy interval of the Rate-Derived Method depends on the metric being measured or calculated and the characteristics of the related transition. IPDV [De02] adds uncertainty to the amount of packets received in a Packet Sampling Interval and this uncertainty adds to the measurement error. The effect of IPDV is not accounted for in the calculation of the accuracy intervals below. IPDV is of importance for the convergence instants were a variation in Forwarding Rate needs to be observed (Convergence Recovery Instant and for topologies with ECMP also Convergence Event Instant and First Route Convergence Instant).

If the Convergence Event Instant is observed on the dataplane using the Rate Derived Method, it needs to be instantaneous for all routes (see Section 4.1). The actual value of the Convergence Event Instant falls within the accuracy interval [-(Packet Sampling Interval + 1/Offered Load), +0] around the value as measured using the Rate-Derived Method.

If the Convergence Recovery Transition is non-instantaneous for all routes then the actual value of the First Route Convergence Instant falls within the accuracy interval [-(Packet Sampling Interval + time between two consecutive packets to the same destination), +0] around the value as measured using the Rate-Derived Method, and the actual value of the Convergence Recovery Instant falls within the accuracy interval [-(2 * Packet Sampling Interval), -(Packet Sampling Interval - time between two consecutive packets to the same destination)] around the value as measured using the Rate-Derived Method.

The term "time between two consecutive packets to the same destination" is added in the above accuracy intervals since packets are sent in a particular order to all destinations in a stream and when part of the routes experience packet loss, it is unknown where in the transmit cycle packets to these routes are sent. This uncertainty adds to the error.

The accuracy intervals of the derived metrics First Route Convergence Time and Rate-Derived Convergence Time are calculated from the above convergence instants accuracy intervals. The actual value of First Route Convergence Time falls within the accuracy interval [-(Packet Sampling Interval + time between two consecutive packets to the same destination), +(Packet Sampling Interval + 1/Offered Load)] around the calculated value. The actual value of Rate-Derived Convergence Time falls within the accuracy interval [-(2 * Packet Sampling Interval), +(time between two consecutive packets to the same destination + 1/Offered Load)] around the calculated value.

6.3. Route-Specific Loss-Derived Method

6.3.1. Tester Capabilities

The Offered Load consists of multiple Streams. The Tester MUST measure packet loss for each Stream separately.

In order to verify Full Convergence completion and the Sustained Convergence Validation Time, the Tester MUST measure packet loss each Packet Sampling Interval. This measurement at each Packet Sampling Interval MAY be per Stream.

Only the total packet loss measured per Stream at the end of the Sustained Convergence Validation Time is used to calculate the benchmark metrics with this method.

6.3.2. Benchmark Metrics

The Route-Specific Loss-Derived Method SHOULD be used to measure Route-Specific Convergence Times. It is the RECOMMENDED method to measure Route Loss of Connectivity Period.

Under the conditions explained in Section 4, First Route Convergence Time and Full Convergence Time as benchmarked using Rate-Derived Method, may be equal to the minimum resp. maximum of the Route-Specific Convergence Times.

6.3.3. Measurement Accuracy

The actual value falls within the accuracy interval [-(number of destinations/Offered Load), +(number of destinations/Offered Load)] around the value as measured using the Route-Specific Loss-Derived Method.

7. Reporting Format

For each test case, it is recommended that the reporting tables below are completed and all time values SHOULD be reported with resolution as specified in [Po10t].

Parameter	Units
Test Case	test case number
Test Topology	Test Topology Figure number
IGP	(ISIS, OSPF, other)
Interface Type	(GigE, POS, ATM, other)
Packet Size offered to DUT	bytes
Offered Load	packets per second
IGP Routes advertised to DUT	number of IGP routes
Nodes in emulated network	number of nodes
Number of Parallel or ECMP links	number of links
Number of Routes measured	number of routes
Packet Sampling Interval on Tester	seconds
Forwarding Delay Threshold	seconds
Timer Values configured on DUT:	
Interface failure indication delay	seconds
IGP Hello Timer	seconds
IGP Dead-Interval or hold-time	seconds
LSA Generation Delay	seconds
LSA Flood Packet Pacing	seconds
LSA Retransmission Packet Pacing	seconds
SPF Delay	seconds

Test Details:

If the Offered Load matches a subset of routes, describe how this subset is selected.

Describe how the Convergence Event is applied; does it cause instantaneous traffic loss or not.

Complete the table below for the initial Convergence Event and the reversion Convergence Event.

```
Parameter                                  Units
------------------------------------------ ----------------------
Conversion Event                           (initial or reversion)

Traffic Forwarding Metrics:
 Total number of packets offered to DUT    number of Packets
 Total number of packets forwarded by DUT  number of Packets
 Connectivity Packet Loss                  number of Packets
 Convergence Packet Loss                   number of Packets
 Out-of-Order Packets                      number of Packets
 Duplicate Packets                         number of Packets

Convergence Benchmarks:
 Rate-Derived Method:
  First Route Convergence Time             seconds
  Full Convergence Time                    seconds
 Loss-Derived Method:
  Loss-Derived Convergence Time            seconds
 Route-Specific Loss-Derived Method:
  Route-Specific Convergence Time[n]       array of seconds
  Minimum R-S Convergence Time             seconds
  Maximum R-S Convergence Time             seconds
  Median R-S Convergence Time              seconds
  Average R-S Convergence Time             seconds

Loss of Connectivity Benchmarks:
 Loss-Derived Method:
  Loss-Derived Loss of Connectivity Period seconds
 Route-Specific Loss-Derived Method:
  Route LoC Period[n]                      array of seconds
  Minimum Route LoC Period                 seconds
  Maximum Route LoC Period                 seconds
  Median Route LoC Period                  seconds
  Average Route LoC Period                 seconds
```

8. Test Cases

It is RECOMMENDED that all applicable test cases be performed for best characterization of the DUT. The test cases follow a generic procedure tailored to the specific DUT configuration and Convergence Event [Po10t]. This generic procedure is as follows:

1. Establish DUT and Tester configurations and advertise an IGP topology from Tester to DUT.

2. Send Offered Load from Tester to DUT on ingress interface.

3. Verify traffic is routed correctly. Verify if traffic is forwarded without drops, without Out-of-Order Packets, and without exceeding the Forwarding Delay Threshold [Po06].

4. Introduce Convergence Event [Po10t].

5. Measure First Route Convergence Time [Po10t].

6. Measure Full Convergence Time [Po10t].

7. Stop Offered Load.

8. Measure Route-Specific Convergence Times, Loss-Derived Convergence Time, Route LoC Periods, and Loss-Derived LoC Period [Po10t].

9. Wait sufficient time for queues to drain. The duration of this time period is equal to the Forwarding Delay Threshold. In absence of a Forwarding Delay Threshold specification the duration of this time period is 2 seconds [Br99].

10. Restart Offered Load.

11. Reverse Convergence Event.

12. Measure First Route Convergence Time.

13. Measure Full Convergence Time.

14. Stop Offered Load.

15. Measure Route-Specific Convergence Times, Loss-Derived Convergence Time, Route LoC Periods, and Loss-Derived LoC Period.

8.1. Interface failures

8.1.1. Convergence Due to Local Interface Failure

Objective

To obtain the IGP convergence times due to a Local Interface failure event. The Next-Best Egress Interface can be a single interface (Figure 1) or an ECMP set (Figure 2). The test with ECMP topology (Figure 2) is OPTIONAL.

Procedure

1. Advertise an IGP topology from Tester to DUT using the topology shown in Figure 1 or 2.

2. Send Offered Load from Tester to DUT on ingress interface.

3. Verify traffic is forwarded over Preferred Egress Interface.

4. Remove link on DUT’s Preferred Egress Interface. This is the Convergence Event.

5. Measure First Route Convergence Time.

6. Measure Full Convergence Time.

7. Stop Offered Load.

8. Measure Route-Specific Convergence Times and Loss-Derived Convergence Time.

9. Wait sufficient time for queues to drain.

10. Restart Offered Load.

11. Restore link on DUT’s Preferred Egress Interface.

12. Measure First Route Convergence Time.

13. Measure Full Convergence Time.

14. Stop Offered Load.

15. Measure Route-Specific Convergence Times, Loss-Derived Convergence Time, Route LoC Periods, and Loss-Derived LoC Period.

Results

The measured IGP convergence time may be influenced by the link failure indication time, LSA/LSP delay, LSA/LSP generation time, LSA/LSP flood packet pacing, SPF delay, SPF execution time, and routing and forwarding tables update time [Po09a].

8.1.2. Convergence Due to Remote Interface Failure

Objective

To obtain the IGP convergence time due to a Remote Interface failure event. The Next-Best Egress Interface can be a single interface

(Figure 3) or an ECMP set (Figure 4). The test with ECMP topology (Figure 4) is OPTIONAL.

Procedure

1. Advertise an IGP topology from Tester to SUT using the topology shown in Figure 3 or 4.

2. Send Offered Load from Tester to SUT on ingress interface.

3. Verify traffic is forwarded over Preferred Egress Interface.

4. Remove link on Tester’s interface [Po10t] connected to SUT’s Preferred Egress Interface. This is the Convergence Event.

5. Measure First Route Convergence Time.

6. Measure Full Convergence Time.

7. Stop Offered Load.

8. Measure Route-Specific Convergence Times and Loss-Derived Convergence Time.

9. Wait sufficient time for queues to drain.

10. Restart Offered Load.

11. Restore link on Tester’s interface connected to DUT’s Preferred Egress Interface.

12. Measure First Route Convergence Time.

13. Measure Full Convergence Time.

14. Stop Offered Load.

15. Measure Route-Specific Convergence Times, Loss-Derived Convergence Time, Route LoC Periods, and Loss-Derived LoC Period.

Results

The measured IGP convergence time may be influenced by the link failure indication time, LSA/LSP delay, LSA/LSP generation time, LSA/ LSP flood packet pacing, SPF delay, SPF execution time, and routing and forwarding tables update time. This test case may produce Stale Forwarding [Po10t] due to a transient microloop between R1 and R2 during convergence, which may increase the measured convergence times and loss of connectivity periods.

8.1.3. Convergence Due to ECMP Member Local Interface Failure

Objective

To obtain the IGP convergence time due to a Local Interface link failure event of an ECMP Member.

Procedure

1. Advertise an IGP topology from Tester to DUT using the test setup shown in Figure 5.

2. Send Offered Load from Tester to DUT on ingress interface.

3. Verify traffic is forwarded over the DUT’s ECMP member interface that will be failed in the next step.

4. Remove link on one of the DUT’s ECMP member interfaces. This is the Convergence Event.

5. Measure First Route Convergence Time.

6. Measure Full Convergence Time.

7. Stop Offered Load.

8. Measure Route-Specific Convergence Times and Loss-Derived Convergence Time. At the same time measure Out-of-Order Packets [Po06] and Duplicate Packets [Po06].

9. Wait sufficient time for queues to drain.

10. Restart Offered Load.

11. Restore link on DUT’s ECMP member interface.

12. Measure First Route Convergence Time.

13. Measure Full Convergence Time.

14. Stop Offered Load.

15. Measure Route-Specific Convergence Times, Loss-Derived Convergence Time, Route LoC Periods, and Loss-Derived LoC Period. At the same time measure Out-of-Order Packets [Po06]

and Duplicate Packets [Po06].

Results

The measured IGP Convergence time may be influenced by link failure indication time, LSA/LSP delay, LSA/LSP generation time, LSA/LSP flood packet pacing, SPF delay, SPF execution time, and routing and forwarding tables update time [Po09a].

8.1.4. Convergence Due to ECMP Member Remote Interface Failure

Objective

To obtain the IGP convergence time due to a Remote Interface link failure event for an ECMP Member.

Procedure

1. Advertise an IGP topology from Tester to DUT using the test setup shown in Figure 6.

2. Send Offered Load from Tester to DUT on ingress interface.

3. Verify traffic is forwarded over the DUT’s ECMP member interface that will be failed in the next step.

4. Remove link on Tester’s interface to R2. This is the Convergence Event Trigger.

5. Measure First Route Convergence Time.

6. Measure Full Convergence Time.

7. Stop Offered Load.

8. Measure Route-Specific Convergence Times and Loss-Derived Convergence Time. At the same time measure Out-of-Order Packets [Po06] and Duplicate Packets [Po06].

9. Wait sufficient time for queues to drain.

10. Restart Offered Load.

11. Restore link on Tester’s interface to R2.

12. Measure First Route Convergence Time.

13. Measure Full Convergence Time.

14. Stop Offered Load.

15. Measure Route-Specific Convergence Times, Loss-Derived Convergence Time, Route LoC Periods, and Loss-Derived LoC Period. At the same time measure Out-of-Order Packets [Po06] and Duplicate Packets [Po06].

Results

The measured IGP convergence time may influenced by the link failure indication time, LSA/LSP delay, LSA/LSP generation time, LSA/LSP flood packet pacing, SPF delay, SPF execution time, and routing and forwarding tables update time. This test case may produce Stale Forwarding [Po10t] due to a transient microloop between R1 and R2 during convergence, which may increase the measured convergence times and loss of connectivity periods.

8.1.5. Convergence Due to Parallel Link Interface Failure

Objective

To obtain the IGP convergence due to a local link failure event for a member of a parallel link. The links can be used for data load balancing

Procedure

1. Advertise an IGP topology from Tester to DUT using the test setup shown in Figure 7.

2. Send Offered Load from Tester to DUT on ingress interface.

3. Verify traffic is forwarded over the parallel link member that will be failed in the next step.

4. Remove link on one of the DUT’s parallel link member interfaces. This is the Convergence Event.

5. Measure First Route Convergence Time.

6. Measure Full Convergence Time.

7. Stop Offered Load.

8. Measure Route-Specific Convergence Times and Loss-Derived Convergence Time. At the same time measure Out-of-Order Packets

[Po06] and Duplicate Packets [Po06].

9. Wait sufficient time for queues to drain.

10. Restart Offered Load.

11. Restore link on DUT’s Parallel Link member interface.

12. Measure First Route Convergence Time.

13. Measure Full Convergence Time.

14. Stop Offered Load.

15. Measure Route-Specific Convergence Times, Loss-Derived Convergence Time, Route LoC Periods, and Loss-Derived LoC Period. At the same time measure Out-of-Order Packets [Po06] and Duplicate Packets [Po06].

Results

The measured IGP convergence time may be influenced by the link failure indication time, LSA/LSP delay, LSA/LSP generation time, LSA/LSP flood packet pacing, SPF delay, SPF execution time, and routing and forwarding tables update time [Po09a].

8.2. Other failures

8.2.1. Convergence Due to Layer 2 Session Loss

Objective

To obtain the IGP convergence time due to a local layer 2 loss.

Procedure

1. Advertise an IGP topology from Tester to DUT using the topology shown in Figure 1.

2. Send Offered Load from Tester to DUT on ingress interface.

3. Verify traffic is routed over Preferred Egress Interface.

4. Remove Layer 2 session from DUT’s Preferred Egress Interface. This is the Convergence Event.

5. Measure First Route Convergence Time.

6. Measure Full Convergence Time.

7. Stop Offered Load.

8. Measure Route-Specific Convergence Times, Loss-Derived Convergence Time, Route LoC Periods, and Loss-Derived LoC Period.

9. Wait sufficient time for queues to drain.

10. Restart Offered Load.

11. Restore Layer 2 session on DUT’s Preferred Egress Interface.

12. Measure First Route Convergence Time.

13. Measure Full Convergence Time.

14. Stop Offered Load.

15. Measure Route-Specific Convergence Times, Loss-Derived Convergence Time, Route LoC Periods, and Loss-Derived LoC Period.

Results

The measured IGP Convergence time may be influenced by the Layer 2 failure indication time, LSA/LSP delay, LSA/LSP generation time, LSA/LSP flood packet pacing, SPF delay, SPF execution time, and routing and forwarding tables update time [Po09a].

Discussion

Configure IGP timers such that the IGP adjacency does not time out before layer 2 failure is detected.

To measure convergence time, traffic SHOULD start dropping on the Preferred Egress Interface on the instant the layer 2 session is removed. Alternatively the Tester SHOULD record the time the instant layer 2 session is removed and traffic loss SHOULD only be measured on the Next-Best Egress Interface. For loss-derived benchmarks the time of the Start Traffic Instant SHOULD be recorded as well. See Section 4.1.

8.2.2. Convergence Due to Loss of IGP Adjacency

Objective

To obtain the IGP convergence time due to loss of an IGP Adjacency.

Procedure

1. Advertise an IGP topology from Tester to DUT using the topology shown in Figure 1.

2. Send Offered Load from Tester to DUT on ingress interface.

3. Verify traffic is routed over Preferred Egress Interface.

4. Remove IGP adjacency from the Preferred Egress Interface while the layer 2 session MUST be maintained. This is the Convergence Event.

5. Measure First Route Convergence Time.

6. Measure Full Convergence Time.

7. Stop Offered Load.

8. Measure Route-Specific Convergence Times, Loss-Derived Convergence Time, Route LoC Periods, and Loss-Derived LoC Period.

9. Wait sufficient time for queues to drain.

10. Restart Offered Load.

11. Restore IGP session on DUT’s Preferred Egress Interface.

12. Measure First Route Convergence Time.

13. Measure Full Convergence Time.

14. Stop Offered Load.

15. Measure Route-Specific Convergence Times, Loss-Derived Convergence Time, Route LoC Periods, and Loss-Derived LoC Period.

Results

The measured IGP Convergence time may be influenced by the IGP Hello Interval, IGP Dead Interval, LSA/LSP delay, LSA/LSP generation time, LSA/LSP flood packet pacing, SPF delay, SPF execution time, and routing and forwarding tables update time [Po09a].

Discussion

Configure layer 2 such that layer 2 does not time out before IGP adjacency failure is detected.

To measure convergence time, traffic SHOULD start dropping on the Preferred Egress Interface on the instant the IGP adjacency is removed. Alternatively the Tester SHOULD record the time the instant the IGP adjacency is removed and traffic loss SHOULD only be measured on the Next-Best Egress Interface. For loss-derived benchmarks the time of the Start Traffic Instant SHOULD be recorded as well. See Section 4.1.

8.2.3. Convergence Due to Route Withdrawal

Objective

To obtain the IGP convergence time due to route withdrawal.

Procedure

1. Advertise an IGP topology from Tester to DUT using the topology shown in Figure 1. The routes that will be withdrawn MUST be a set of leaf routes advertised by at least two nodes in the emulated topology. The topology SHOULD be such that before the withdrawal the DUT prefers the leaf routes advertised by a node "nodeA" via the Preferred Egress Interface, and after the withdrawal the DUT prefers the leaf routes advertised by a node "nodeB" via the Next-Best Egress Interface.

2. Send Offered Load from Tester to DUT on Ingress Interface.

3. Verify traffic is routed over Preferred Egress Interface.

4. The Tester withdraws the set of IGP leaf routes from nodeA. This is the Convergence Event. The withdrawal update message SHOULD be a single unfragmented packet. If the routes cannot be withdrawn by a single packet, the messages SHOULD be sent using the same pacing characteristics as the DUT. The Tester MAY record the time it sends the withdrawal message(s).

5. Measure First Route Convergence Time.

6. Measure Full Convergence Time.

7. Stop Offered Load.

8. Measure Route-Specific Convergence Times, Loss-Derived Convergence Time, Route LoC Periods, and Loss-Derived LoC Period.

9. Wait sufficient time for queues to drain.

10. Restart Offered Load.

11. Re-advertise the set of withdrawn IGP leaf routes from nodeA emulated by the Tester. The update message SHOULD be a single unfragmented packet. If the routes cannot be advertised by a single packet, the messages SHOULD be sent using the same pacing characteristics as the DUT. The Tester MAY record the time it sends the update message(s).

12. Measure First Route Convergence Time.

13. Measure Full Convergence Time.

14. Stop Offered Load.

15. Measure Route-Specific Convergence Times, Loss-Derived Convergence Time, Route LoC Periods, and Loss-Derived LoC Period.

Results

The measured IGP convergence time is influenced by SPF or route calculation delay, SPF or route calculation execution time, and routing and forwarding tables update time [Po09a].

Discussion

To measure convergence time, traffic SHOULD start dropping on the Preferred Egress Interface on the instant the routes are withdrawn by the Tester. Alternatively the Tester SHOULD record the time the instant the routes are withdrawn and traffic loss SHOULD only be measured on the Next-Best Egress Interface. For loss-derived benchmarks the time of the Start Traffic Instant SHOULD be recorded as well. See Section 4.1.

8.3. Administrative changes

8.3.1. Convergence Due to Local Adminstrative Shutdown

Objective

To obtain the IGP convergence time due to taking the DUT's Local

Interface administratively out of service.

Procedure

1. Advertise an IGP topology from Tester to DUT using the topology shown in Figure 1.

2. Send Offered Load from Tester to DUT on ingress interface.

3. Verify traffic is routed over Preferred Egress Interface.

4. Take the DUT's Preferred Egress Interface administratively out of service. This is the Convergence Event.

5. Measure First Route Convergence Time.

6. Measure Full Convergence Time.

7. Stop Offered Load.

8. Measure Route-Specific Convergence Times, Loss-Derived Convergence Time, Route LoC Periods, and Loss-Derived LoC Period.

9. Wait sufficient time for queues to drain.

10. Restart Offered Load.

11. Restore Preferred Egress Interface by administratively enabling the interface.

12. Measure First Route Convergence Time.

13. Measure Full Convergence Time.

14. Stop Offered Load.

15. Measure Route-Specific Convergence Times, Loss-Derived Convergence Time, Route LoC Periods, and Loss-Derived LoC Period.

16. It is possible that no measured packet loss will be observed for this test case.

Results

The measured IGP Convergence time may be influenced by LSA/LSP delay, LSA/LSP generation time, LSA/LSP flood packet pacing, SPF delay, SPF execution time, and routing and forwarding tables update time [Po09a].

8.3.2. Convergence Due to Cost Change

Objective

To obtain the IGP convergence time due to route cost change.

Procedure

1. Advertise an IGP topology from Tester to DUT using the topology shown in Figure 1.

2. Send Offered Load from Tester to DUT on ingress interface.

3. Verify traffic is routed over Preferred Egress Interface.

4. The Tester, emulating the neighbor node, increases the cost for all IGP routes at DUT's Preferred Egress Interface so that the Next-Best Egress Interface becomes preferred path. The update message advertising the higher cost MUST be a single unfragmented packet. This is the Convergence Event. The Tester MAY record the time it sends the update message advertising the higher cost on the Preferred Egress Interface.

5. Measure First Route Convergence Time.

6. Measure Full Convergence Time.

7. Stop Offered Load.

8. Measure Route-Specific Convergence Times, Loss-Derived Convergence Time, Route LoC Periods, and Loss-Derived LoC Period.

9. Wait sufficient time for queues to drain.

10. Restart Offered Load.

11. The Tester, emulating the neighbor node, decreases the cost for all IGP routes at DUT's Preferred Egress Interface so that the Preferred Egress Interface becomes preferred path. The update message advertising the lower cost MUST be a single unfragmented packet.

12. Measure First Route Convergence Time.

13. Measure Full Convergence Time.

14. Stop Offered Load.

15. Measure Route-Specific Convergence Times, Loss-Derived Convergence Time, Route LoC Periods, and Loss-Derived LoC Period.

Results

The measured IGP Convergence time may be influenced by SPF delay, SPF execution time, and routing and forwarding tables update time [Po09a].

Discussion

To measure convergence time, traffic SHOULD start dropping on the Preferred Egress Interface on the instant the cost is changed by the Tester. Alternatively the Tester SHOULD record the time the instant the cost is changed and traffic loss SHOULD only be measured on the Next-Best Egress Interface. For loss-derived benchmarks the time of the Start Traffic Instant SHOULD be recorded as well. See Section 4.1.

9. 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.

10. IANA Considerations

This document requires no IANA considerations.

11. Acknowledgements

Thanks to Sue Hares, Al Morton, Kevin Dubray, Ron Bonica, David Ward, Peter De Vriendt, Anuj Dewagan and the BMWG for their contributions to this work.

12. Normative References

[Br91] Bradner, S., "Benchmarking terminology for network interconnection devices", RFC 1242, July 1991.

[Br97] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

[Br99] Bradner, S. and J. McQuaid, "Benchmarking Methodology for Network Interconnect Devices", RFC 2544, March 1999.

[Ca90] Callon, R., "Use of OSI IS-IS for routing in TCP/IP and dual environments", RFC 1195, December 1990.

[Co08] Coltun, R., Ferguson, D., Moy, J., and A. Lindem, "OSPF for IPv6", RFC 5340, July 2008.

[De02] Demichelis, C. and P. Chimento, "IP Packet Delay Variation Metric for IP Performance Metrics (IPPM)", RFC 3393, November 2002.

[Ho08] Hopps, C., "Routing IPv6 with IS-IS", RFC 5308, October 2008.

[Ko02] Koodli, R. and R. Ravikanth, "One-way Loss Pattern Sample Metrics", RFC 3357, August 2002.

[Ma98] Mandeville, R., "Benchmarking Terminology for LAN Switching Devices", RFC 2285, February 1998.

[Mo98] Moy, J., "OSPF Version 2", STD 54, RFC 2328, April 1998.

[Po06] Poretsky, S., Perser, J., Erramilli, S., and S. Khurana, "Terminology for Benchmarking Network-layer Traffic Control Mechanisms", RFC 4689, October 2006.

[Po09a] Poretsky, S., "Considerations for Benchmarking Link-State IGP Data Plane Route Convergence", draft-ietf-bmwg-igp-dataplane-conv-app-17 (work in progress), March 2009.

[Po10t] Poretsky, S., Imhoff, B., and K. Michielsen, "Terminology for Benchmarking Link-State IGP Data Plane Route Convergence", draft-ietf-bmwg-igp-dataplane-conv-term-20 (work in progress), March 2010.

Authors' Addresses

Scott Poretsky
Allot Communications
67 South Bedford Street, Suite 400
Burlington, MA 01803
USA

Phone: + 1 508 309 2179
Email: sporetsky@allot.com

Brent Imhoff
Juniper Networks
1194 North Mathilda Ave
Sunnyvale, CA 94089
USA

Phone: + 1 314 378 2571
Email: bimhoff@planetspork.com

Kris Michielsen
Cisco Systems
6A De Kleetlaan
Diegem, BRABANT 1831
Belgium

Email: kmichiel@cisco.com

Network Working Group
Internet-Draft
Intended status: Informational
Expires: May 12, 2011

S. Poretsky
Allot Communications
B. Imhoff
Juniper Networks
K. Michielsen
Cisco Systems

November 8, 2010

Terminology for Benchmarking Link-State IGP Data Plane Route Convergence
draft-ietf-bmwg-igp-dataplane-conv-term-22

Abstract

This document describes the terminology for benchmarking Interior Gateway Protocol (IGP) Route Convergence. The terminology is to be used for benchmarking IGP convergence time through externally observable (black box) data plane measurements. The terminology can be applied to any link-state IGP, such as ISIS and OSPF.

Status of this Memo

Copyright Notice

Table of Contents

1. Introduction and Scope

This draft describes the terminology for benchmarking Link-State Interior Gateway Protocol (IGP) Convergence. The motivation and applicability for this benchmarking is provided in [Po09a]. The methodology to be used for this benchmarking is described in [Po10m]. The purpose of this document is to introduce new terms required to complete execution of the IGP Route Methodology [Po10m].

IGP convergence time is measured on the data plane at the Tester by observing packet loss through the DUT. The methodology and terminology to be used for benchmarking IGP Convergence can be applied to IPv4 and IPv6 traffic and link-state IGPs such as ISIS [Ca90][Ho08], OSPF [Mo98][Co08], and others.

2. Existing Definitions

This document uses existing terminology defined in other BMWG work. Examples include, but are not limited to:

Frame Loss Rate	[Ref.[Br91], section 3.6]
Throughput	[Ref.[Br91], section 3.17]
Offered Load	[Ref.[Ma98], section 3.5.2]
Forwarding Rate	[Ref.[Ma98], section 3.6.1]
Device Under Test (DUT)	[Ref.[Ma98], section 3.1.1]
System Under Test (SUT)	[Ref.[Ma98], section 3.1.2]
Out-of-Order Packet	[Ref.[Po06], section 3.3.4]
Duplicate Packet	[Ref.[Po06], section 3.3.5]
Packet Reordering	[Ref.[Mo06], section 3.3]
Stream	[Ref.[Po06], section 3.3.2]
Forwarding Delay	[Ref.[Po06], section 3.2.4]
IP Packet Delay Variation (IPDV)	[Ref.[De02], section 1.2]
Loss Period	[Ref.[Ko02], section 4]

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 BCP 14, RFC 2119 [Br97]. RFC 2119 defines the use of these key words to help make the intent of standards track documents as clear as possible. While this document uses these keywords, this document is not a standards track document.

3. Term Definitions

3.1. Convergence Types

3.1.1. Route Convergence

Definition:

The process of updating all components of the router, including the Routing Information Base (RIB) and Forwarding Information Base (FIB), along with software and hardware tables, with the most recent route change(s) such that forwarding for a route entry is successful on the Next-Best Egress Interface.

Discussion:

Route Convergence MUST occur after a Convergence Event. Route Convergence can be observed externally by the rerouting of data traffic for a destination matching a route entry to the Next-best Egress Interface. Completion of Route Convergence may or may not be sustained over time.

Measurement Units: N/A

Issues: None

See Also:

Network Convergence, Full Convergence, Convergence Event

3.1.2. Full Convergence

Definition:

Route Convergence for all routes in the FIB.

Discussion:

Full Convergence MUST occur after a Convergence Event. Full Convergence can be observed externally by the rerouting of data traffic to destinations matching all route entries to the Next-best Egress Interface. Completion of Full Convergence is externally observable from the data plane when the Forwarding Rate of the data plane traffic on the Next-Best Egress Interface equals the Offered Load.

Completion of Full Convergence may or may not be sustained over time.

Measurement Units: N/A

Issues: None

See Also:

Network Convergence, Route Convergence, Convergence Event, Full Convergence Time, Convergence Recovery Instant

3.1.3. Network Convergence

Definition:

Full Convergence in all routers throughout the network.

Discussion:

Network Convergence includes all Route Convergence operations for all routers in the network following a Convergence Event.

Completion of Network Convergence can be observed by recovery of the network Forwarding Rate to equal the Offered Load, with no Stale Forwarding, and no Blenders [Ca01][Ci03].

Completion of Network Convergence may or may not be sustained over time.

Measurement Units: N/A

Issues: None

See Also:

Route Convergence, Full Convergence, Stale Forwarding

3.2. Instants

3.2.1. Traffic Start Instant

Definition:

The time instant the Tester sends out the first data packet to the DUT.

Discussion:

If using the Loss-Derived Method or the Route-Specific Loss-Derived Method to benchmark IGP convergence time, and the applied Convergence Event does not cause instantaneous traffic loss for all routes at the Convergence Event Instant then the Tester SHOULD collect a timestamp on the Traffic Start Instant in order to measure the period of time between the Traffic Start Instant and Convergence Event Instant.

Measurement Units:

hh:mm:ss:nnn:uuu, where 'nnn' is milliseconds and 'uuu' is microseconds.

Issues: None

See Also:

Convergence Event Instant, Route-Specific Convergence Time, Loss-Derived Convergence Time.

3.2.2. Convergence Event Instant

Definition:

The time instant that a Convergence Event occurs.

Discussion:

If the Convergence Event causes instantaneous traffic loss on the Preferred Egress Interface, the Convergence Event Instant is observable from the data plane as the instant that the DUT begins to exhibit packet loss.

The Tester SHOULD collect a timestamp on the Convergence Event Instant if it is not observable from the data plane.

Measurement Units:

hh:mm:ss:nnn:uuu, where 'nnn' is milliseconds and 'uuu' is microseconds.

Issues: None

See Also: Convergence Event

3.2.3. Convergence Recovery Instant

Definition:

The time instant that Full Convergence has completed.

Discussion:

The Full Convergence completed state MUST be maintained for an interval of duration equal to the Sustained Convergence Validation Time in order to validate the Convergence Recovery Instant.

The Convergence Recovery Instant is observable from the data plane as the instant the DUT forwards traffic to all destinations over the Next-Best Egress Interface.

Measurement Units:

hh:mm:ss:nnn:uuu, where 'nnn' is milliseconds and 'uuu' is microseconds.

Issues: None

See Also:

Sustained Convergence Validation Time, Full Convergence

3.2.4. First Route Convergence Instant

Definition:

The time instant the first route entry completes Route Convergence following a Convergence Event

Discussion:

Any route may be the first to complete Route Convergence. The First Route Convergence Instant is observable from the data plane as the instant that the first packet is received from the Next-Best Egress Interface.

Measurement Units:

hh:mm:ss:nnn:uuu, where 'nnn' is milliseconds and 'uuu' is microseconds.

Issues: None

See Also: Route Convergence

3.3. Transitions

3.3.1. Convergence Event Transition

Definition:

A time interval following a Convergence Event in which Forwarding Rate on the Preferred Egress Interface gradually reduces to zero.

Discussion:

The Forwarding Rate during a Convergence Event Transition may not decrease linearly.

The Forwarding Rate observed on all DUT egress interfaces may or may not decrease to zero.

The Offered Load, the number of routes, and the Packet Sampling Interval influence the observations of the Convergence Event Transition using the Rate-Derived Method. This is further discussed with the term "Rate-Derived Method".

Measurement Units: seconds

Issues: None

See Also:

Convergence Event, Rate-Derived Method

3.3.2. Convergence Recovery Transition

Definition:

A time interval following the First Route Convergence Instant in which Forwarding Rate on the Next-Best Egress Interface gradually increases to equal the Offered Load.

Discussion:

The Forwarding Rate observed during a Convergence Recovery Transition may not increase linearly.

The Offered Load, the number of routes, and the Packet Sampling Interval influence the observations of the Convergence Recovery Transition using the Rate-Derived Method. This is further discussed with the term "Rate-Derived Method".

Measurement Units: seconds

Issues: None

See Also:

Full Convergence,First Route Convergence Instant, Rate-Derived Method

3.4. Interfaces

3.4.1. Local Interface

Definition:

An interface on the DUT.

Discussion:

A failure of the Local Interface indicates that the failure occurred directly on the DUT.

Measurement Units: N/A

Issues: None

See Also: Remote Interface

3.4.2. Remote Interface

Definition:

An interface on a neighboring router that is not directly connected to any interface on the DUT.

Discussion:

A failure of a Remote Interface indicates that the failure occurred on a neighbor router’s interface that is not directly connected to the DUT.

Measurement Units: N/A

Issues: None

See Also: Local Interface

3.4.3. Preferred Egress Interface

Definition:

The outbound interface from the DUT for traffic routed to the preferred next-hop.

Discussion:

The Preferred Egress Interface is the egress interface prior to a Convergence Event.

Measurement Units: N/A

Issues: None

See Also: Next-Best Egress Interface

3.4.4. Next-Best Egress Interface

Definition:

The outbound interface from the DUT for traffic routed to the second-best next-hop.

Discussion:

The Next-Best Egress Interface becomes the egress interface after a Convergence Event.

Measurement Units: N/A

Issues: None

See Also: Preferred Egress Interface

3.5. Benchmarking Methods

3.5.1. Rate-Derived Method

Definition:

The method to calculate convergence time benchmarks from observing Forwarding Rate each Packet Sampling Interval.

Discussion:

Figure 1 shows an example of the Forwarding Rate change in time during convergence as observed when using the Rate-Derived Method.

```
           ^         Traffic                      Convergence
      Fwd  |         Start                        Recovery
      Rate |         Instant                      Instant
           | Offered  ^                             ^
           | Load --> ----------\                   /-----------
           |                     \                 /<--- Convergence
           |                      \     Packet    /      Recovery
           |       Convergence --->\     Loss    /       Transition
           |       Event            \           /
           |       Transition        \---------/ <-- Max Packet Loss
           |
           +---------------------------------------------------------->
                           ^                   ^                 time
                      Convergence         First Route
                      Event Instant       Convergence Instant
```

Figure 1: Rate-Derived Convergence Graph

The Offered Load SHOULD consist of a single Stream [Po06]. If sending multiple Streams, the measured traffic rate statistics for all Streams MUST be added together.

The destination addresses for the Offered Load MUST be distributed such that all routes or a statistically representative subset of all routes are matched and each of these routes is offered an equal share of the Offered Load. It is RECOMMENDED to send traffic to all routes, but a statistically representative subset of all routes can be used if required.

At least one packet per route for all routes matched in the Offered Load MUST be offered to the DUT within each Packet Sampling Interval. For maximum accuracy the value for the Packet Sampling Interval SHOULD be as small as possible, but the presence of IP Packet Delay Variation (IPDV) [De02] may enforce using a larger Packet Sampling Interval.

The Offered Load, IPDV, the number of routes, and the Packet Sampling Interval influence the observations for the Rate-Derived Method. It may be difficult to identify the different convergence time instants in the Rate-Derived Convergence Graph. For example, it is possible that a Convergence Event causes the Forwarding Rate to drop to zero, while this may not be observed in the Forwarding Rate measurements if the Packet Sampling Interval is too large.

IPDV causes fluctuations in the number of received packets during each Packet Sampling Interval. To account for the presence of IPDV in determining if a convergence instant has been reached, Forwarding Delay SHOULD be observed during each Packet Sampling Interval. The minimum and maximum number of packets expected in a Packet Sampling Interval in presence of IPDV can be calculated with Equation 1.

```
number of packets expected in a Packet Sampling Interval
  in presence of IP Packet Delay Variation
    = expected number of packets without IP Packet Delay Variation
      +/-( (maxDelay - minDelay) * Offered Load)
with minDelay and maxDelay the minimum resp. maximum Forwarding Delay
  of packets received during the Packet Sampling Interval
```

Equation 1

To determine if a convergence instant has been reached the number of packets received in a Packet Sampling Interval is compared with the range of expected number of packets calculated in Equation 1.

If packets are going over multiple ECMP members and one or more of the members has failed then the number of received packets during each Packet Sampling Interval may vary, even excluding presence of IPDV. To prevent fluctuation of the number of received packets during each Packet Sampling Interval for this reason, the Packet Sampling Interval duration SHOULD be a whole multiple of the time between two consecutive packets sent to the same destination.

Metrics measured at the Packet Sampling Interval MUST include Forwarding Rate and packet loss.

Rate-Derived Method is a RECOMMENDED method to measure convergence time benchmarks.

To measure convergence time benchmarks for Convergence Events that do not cause instantaneous traffic loss for all routes at the Convergence Event Instant, the Tester SHOULD collect a timestamp of the Convergence Event Instant and the Tester SHOULD observe Forwarding Rate separately on the Next-Best Egress Interface.

Since the Rate-Derived Method does not distinguish between individual traffic destinations, it SHOULD NOT be used for any route specific measurements. Therefor Rate-Derived Method SHOULD NOT be used to benchmark Route Loss of Connectivity Period.

Measurement Units: N/A

Issues: None

See Also:

Packet Sampling Interval, Convergence Event, Convergence Event

Instant, Full Convergence

3.5.2. Loss-Derived Method

Definition:

The method to calculate the Loss-Derived Convergence Time and Loss-Derived Loss of Connectivity Period benchmarks from the amount of packet loss.

Discussion:

The Offered Load SHOULD consist of a single Stream [Po06]. If sending multiple Streams, the measured traffic rate statistics for all Streams MUST be added together.

The destination addresses for the Offered Load MUST be distributed such that all routes or a statistically representative subset of all routes are matched and each of these routes is offered an equal share of the Offered Load. It is RECOMMENDED to send traffic to all routes, but a statistically representative subset of all routes can be used if required.

Loss-Derived Method SHOULD always be combined with Rate-Derived Method in order to observe Full Convergence completion. The total amount of Convergence Packet Loss is collected after Full Convergence completion.

To measure convergence time and loss of connectivity benchmarks, the Tester SHOULD in general observe packet loss on all DUT egress interfaces (Connectivity Packet Loss).

To measure convergence time benchmarks for Convergence Events that do not cause instantaneous traffic loss for all routes at the Convergence Event Instant, the Tester SHOULD collect timestamps of the Start Traffic Instant and of the Convergence Event Instant, and the Tester SHOULD observe packet loss separately on the Next-Best Egress Interface (Convergence Packet Loss).

Since Loss-Derived Method does not distinguish between traffic destinations and the packet loss statistics are only collected after Full Convergence completion, this method can only be used to measure average values over all routes. For these reasons Loss-Derived Method can only be used to benchmark Loss-Derived Convergence Time and Loss-Derived Loss of Connectivity Period.

Note that the Loss-Derived Method measures an average over all routes, including the routes that may not be impacted by the

Convergence Event, such as routes via non-impacted members of ECMP or parallel links.

Measurement Units: seconds

Issues: None

See Also:

Loss-Derived Convergence Time, Loss-Derived Loss of Connectivity Period, Convergence Packet Loss

3.5.3. Route-Specific Loss-Derived Method

Definition:

The method to calculate the Route-Specific Convergence Time benchmark from the amount of packet loss during convergence for a specific route entry.

Discussion:

To benchmark Route-Specific Convergence Time, the Tester provides an Offered Load that consists of multiple Streams [Po06]. Each Stream has a single destination address matching a different route entry, for all routes or a statistically representative subset of all routes. Convergence Packet Loss is measured for each Stream separately.

Route-Specific Loss-Derived Method SHOULD always be combined with Rate-Derived Method in order to observe Full Convergence completion. The total amount of Convergence Packet Loss for each Stream is collected after Full Convergence completion.

Route-Specific Loss-Derived Method is a RECOMMENDED method to measure convergence time benchmarks.

To measure convergence time and loss of connectivity benchmarks, the Tester SHOULD in general observe packet loss on all DUT egress interfaces (Connectivity Packet Loss).

To measure convergence time benchmarks for Convergence Events that do not cause instantaneous traffic loss for all routes at the Convergence Event Instant, the Tester SHOULD collect timestamps of the Start Traffic Instant and of the Convergence Event Instant, and the Tester SHOULD observe packet loss separately on the Next-Best Egress Interface (Convergence Packet Loss).

Since Route-Specific Loss-Derived Method uses traffic streams to individual routes, it measures packet loss as it would be experienced by a network user. For this reason Route-Specific Loss-Derived Method is RECOMMENDED to measure Route-Specific Convergence Time benchmarks and Route Loss of Connectivity Period benchmarks.

Measurement Units: seconds

Issues: None

See Also:

Route-Specific Convergence Time, Route Loss of Connectivity Period, Convergence Packet Loss

3.6. Benchmarks

3.6.1. Full Convergence Time

Definition:

The time duration of the period between the Convergence Event Instant and the Convergence Recovery Instant as observed using the Rate-Derived Method.

Discussion:

Using the Rate-Derived Method, Full Convergence Time can be calculated as the time difference between the Convergence Event Instant and the Convergence Recovery Instant, as shown in Equation 2.

```
     Full Convergence Time =
         Convergence Recovery Instant - Convergence Event Instant
```

Equation 2

The Convergence Event Instant can be derived from the Forwarding Rate observation or from a timestamp collected by the Tester.

For the testcases described in [Po10m], it is expected that Full Convergence Time equals the maximum Route-Specific Convergence Time when benchmarking all routes in FIB using the Route-Specific Loss-Derived Method.

It is not possible to measure Full Convergence Time using the Loss-Derived Method.

Measurement Units: seconds

Issues: None

See Also:

Full Convergence, Rate-Derived Method, Route-Specific Loss-Derived Method

3.6.2. First Route Convergence Time

Definition:

The duration of the period between the Convergence Event Instant and the First Route Convergence Instant as observed using the Rate-Derived Method.

Discussion:

Using the Rate-Derived Method, First Route Convergence Time can be calculated as the time difference between the Convergence Event Instant and the First Route Convergence Instant, as shown with Equation 3.

```
First Route Convergence Time =
    First Route Convergence Instant - Convergence Event Instant
```

Equation 3

The Convergence Event Instant can be derived from the Forwarding Rate observation or from a timestamp collected by the Tester.

For the testcases described in [Po10m], it is expected that First Route Convergence Time equals the minimum Route-Specific Convergence Time when benchmarking all routes in FIB using the Route-Specific Loss-Derived Method.

It is not possible to measure First Route Convergence Time using the Loss-Derived Method.

Measurement Units: seconds

Issues: None

See Also:

Rate-Derived Method, Route-Specific Loss-Derived Method, First Route Convergence Instant

3.6.3. Route-Specific Convergence Time

Definition:

The amount of time it takes for Route Convergence to be completed for a specific route, as calculated from the amount of packet loss during convergence for a single route entry.

Discussion:

Route-Specific Convergence Time can only be measured using the Route-Specific Loss-Derived Method.

If the applied Convergence Event causes instantaneous traffic loss for all routes at the Convergence Event Instant, Connectivity Packet Loss should be observed. Connectivity Packet Loss is the combined packet loss observed on Preferred Egress Interface and Next-Best Egress Interface. When benchmarking Route-Specific Convergence Time, Connectivity Packet Loss is measured and Equation 4 is applied for each measured route. The calculation is equal to Equation 8 in Section 3.6.5.

```
Route-Specific Convergence Time =
   Connectivity Packet Loss for specific route/Offered Load per route
```

Equation 4

If the applied Convergence Event does not cause instantaneous traffic loss for all routes at the Convergence Event Instant, then the Tester SHOULD collect timestamps of the Traffic Start Instant and of the Convergence Event Instant, and the Tester SHOULD observe Convergence Packet Loss separately on the Next-Best Egress Interface. When benchmarking Route-Specific Convergence Time, Convergence Packet Loss is measured and Equation 5 is applied for each measured route.

```
Route-Specific Convergence Time =
    Convergence Packet Loss for specific route/Offered Load per route
    - (Convergence Event Instant - Traffic Start Instant)
```

Equation 5

The Convergence Event Instant and Traffic Start Instant SHOULD be collected by the Tester.

The Route-Specific Convergence Time benchmarks enable minimum, maximum, average, and median convergence time measurements to be reported by comparing the results for the different route entries. It also enables benchmarking of convergence time when configuring a priority value for route entry(ies). Since multiple Route-Specific Convergence Times can be measured it is possible to have an array of results. The format for reporting Route-Specific Convergence Time is provided in [Po10m].

Measurement Units: seconds

Issues: None

See Also:

Convergence Event, Convergence Packet Loss, Connectivity Packet Loss, Route Convergence

3.6.4. Loss-Derived Convergence Time

Definition:

The average Route Convergence time for all routes in FIB, as calculated from the amount of packet loss during convergence.

Discussion:

Loss-Derived Convergence Time is measured using the Loss-Derived Method.

If the applied Convergence Event causes instantaneous traffic loss for all routes at the Convergence Event Instant, Connectivity Packet Loss should be observed. Connectivity Packet Loss is the combined packet loss observed on Preferred Egress Interface and Next-Best Egress Interface. When benchmarking Loss-Derived Convergence Time, Connectivity Packet Loss is measured and Equation 6 is applied.

```
              Loss-Derived Convergence Time =
                  Connectivity Packet Loss/Offered Load

                                Equation 6
```

If the applied Convergence Event does not cause instantaneous traffic loss for all routes at the Convergence Event Instant, then the Tester SHOULD collect timestamps of the Start Traffic Instant and of the Convergence Event Instant and the Tester SHOULD observe Convergence Packet Loss separately on the Next-Best Egress Interface. When benchmarking Loss-Derived Convergence Time, Convergence Packet Loss is measured and Equation 7 is applied.

```
Loss-Derived Convergence Time =
    Convergence Packet Loss/Offered Load
    - (Convergence Event Instant - Traffic Start Instant)
```

Equation 7

The Convergence Event Instant and Traffic Start Instant SHOULD be collected by the Tester.

Measurement Units: seconds

Issues: None

See Also:

Convergence Packet Loss, Connectivity Packet Loss, Route Convergence

3.6.5. Route Loss of Connectivity Period

Definition:

The time duration of traffic loss for a specific route entry following a Convergence Event until Full Convergence completion, as observed using the Route-Specific Loss-Derived Method.

Discussion:

In general the Route Loss of Connectivity Period is not equal to the Route-Specific Convergence Time. If the DUT continues to forward traffic to the Preferred Egress Interface after the Convergence Event is applied then the Route Loss of Connectivity Period will be smaller than the Route-Specific Convergence Time. This is also specifically the case after reversing a failure event.

The Route Loss of Connectivity Period may be equal to the Route-Specific Convergence Time if, as a characteristic of the Convergence Event, traffic for all routes starts dropping instantaneously on the Convergence Event Instant. See discussion in [Po10m].

For the testcases described in [Po10m] the Route Loss of Connectivity Period is expected to be a single Loss Period [Ko02].

When benchmarking Route Loss of Connectivity Period, Connectivity Packet Loss is measured for each route and Equation 8 is applied for each measured route entry. The calculation is equal to Equation 4 in Section 3.6.3.

Route Loss of Connectivity Period =
   Connectivity Packet Loss for specific route/Offered Load per route

Equation 8

Route Loss of Connectivity Period SHOULD be measured using Route-Specific Loss-Derived Method.

Measurement Units: seconds

Issues: None

See Also:

Route-Specific Convergence Time, Route-Specific Loss-Derived Method, Connectivity Packet Loss

3.6.6. Loss-Derived Loss of Connectivity Period

Definition:

The average time duration of traffic loss for all routes following a Convergence Event until Full Convergence completion, as observed using the Loss-Derived Method.

Discussion:

In general the Loss-Derived Loss of Connectivity Period is not equal to the Loss-Derived Convergence Time. If the DUT continues to forward traffic to the Preferred Egress Interface after the Convergence Event is applied then the Loss-Derived Loss of Connectivity Period will be smaller than the Loss-Derived Convergence Time. This is also specifically the case after reversing a failure event.

The Loss-Derived Loss of Connectivity Period may be equal to the Loss-Derived Convergence Time if, as a characteristic of the Convergence Event, traffic for all routes starts dropping instantaneously on the Convergence Event Instant. See discussion in [Po10m].

For the testcases described in [Po10m] each route's Route Loss of Connectivity Period is expected to be a single Loss Period [Ko02].

When benchmarking Loss-Derived Loss of Connectivity Period, Connectivity Packet Loss is measured for all routes and Equation 9 is applied. The calculation is equal to Equation 6 in Section 3.6.4.

Loss-Derived Loss of Connectivity Period =
Connectivity Packet Loss for all routes/Offered Load

Equation 9

Loss-Derived Loss of Connectivity Period SHOULD be measured using Loss-Derived Method.

Measurement Units: seconds

Issues: None

See Also:

Loss-Derived Convergence Time, Loss-Derived Method, Connectivity Packet Loss

3.7. Measurement Terms

3.7.1. Convergence Event

Definition:

The occurrence of a planned or unplanned event in the network that will result in a change in the egress interface of the Device Under Test (DUT) for routed packets.

Discussion:

Convergence Events include but are not limited to link loss, routing protocol session loss, router failure, configuration change, and better next-hop learned via a routing protocol.

Measurement Units: N/A

Issues: None

See Also: Convergence Event Instant

3.7.2. Packet Loss

Definition:

The number of packets that should have been forwarded by a DUT under a constant Offered Load that were not forwarded due to lack of resources.

Discussion:

Packet Loss is a modified version of the term "Frame Loss Rate" as defined in [Br91]. The term "Frame Loss" is intended for Ethernet Frames while "Packet Loss" is intended for IP packets.

Measurement units: Number of offered packets that are not forwarded.

Issues: None

See Also: Convergence Packet Loss

3.7.3. Convergence Packet Loss

Definition:

The number of packets lost due to a Convergence Event until Full Convergence completes, as observed on the Next-Best Egress Interface.

Discussion:

Convergence Packet Loss is observed on the Next-Best Egress Interface. It only needs to be observed for Convergence Events that do not cause instantaneous traffic loss at Convergence Event Instant.

Convergence Packet Loss includes packets that were lost and packets that were delayed due to buffering. The maximum acceptable Forwarding Delay (Forwarding Delay Threshold) is a parameter of the methodology, if it is applied it MUST be reported.

Measurement Units: number of packets

Issues: None

See Also:

Packet Loss, Full Convergence, Convergence Event, Connectivity Packet Loss

3.7.4. Connectivity Packet Loss

Definition:

The number of packets lost due to a Convergence Event until Full Convergence completes.

Discussion:

Connectivity Packet Loss is observed on all DUT egress interfaces.

Connectivity Packet Loss includes packets that were lost and packets that were delayed due to buffering. The maximum acceptable Forwarding Delay (Forwarding Delay Threshold) is a parameter of the methodology, if it is applied it MUST be reported.

Measurement Units: number of packets

Issues: None

See Also:

Packet Loss, Route Loss of Connectivity Period, Convergence Event, Convergence Packet Loss

3.7.5. Packet Sampling Interval

Definition:

The interval at which the Tester (test equipment) polls to make measurements for arriving packets.

Discussion:

At least one packet per route for all routes matched in the Offered Load MUST be offered to the DUT within the Packet Sampling Interval. Metrics measured at the Packet Sampling Interval MUST include Forwarding Rate and received packets.

Packet Sampling Interval can influence the convergence graph as observed with the Rate-Derived Method. This is particularly true when implementations complete Full Convergence in less time than the Packet Sampling Interval. The Convergence Event Instant and First Route Convergence Instant may not be easily identifiable and the Rate-Derived Method may produce a larger than actual convergence time.

Using a small Packet Sampling Interval in the presence of IPDV [De02] may cause fluctuations of the Forwarding Rate observation and can prevent correct observation of the different convergence time instants.

The value of the Packet Sampling Interval only contributes to the measurement accuracy of the Rate-Derived Method. For maximum accuracy the value for the Packet Sampling Interval SHOULD be as small as possible, but the presence of IPDV may enforce using a larger Packet Sampling Interval.

Measurement Units: seconds

Issues: None

See Also: Rate-Derived Method

3.7.6. Sustained Convergence Validation Time

Definition:

The amount of time for which the completion of Full Convergence is maintained without additional packet loss.

Discussion:

The purpose of the Sustained Convergence Validation Time is to produce convergence benchmarks protected against fluctuation in Forwarding Rate after the completion of Full Convergence is observed. The RECOMMENDED Sustained Convergence Validation Time to be used is the time to send 5 consecutive packets to each destination with a minimum of 5 seconds. The BMWG selected 5 seconds based upon [Br99] which recommends waiting 2 seconds for residual frames to arrive (this is the Forwarding Delay Threshold for the last packet sent) and 5 seconds for DUT restabilization.

Measurement Units: seconds

Issues: None

See Also:

Full Convergence, Convergence Recovery Instant

3.7.7. Forwarding Delay Threshold

Definition:

The maximum waiting time threshold used to distinguish between packets with very long delay and lost packets that will never arrive.

Discussion:

Applying a Forwarding Delay Threshold allows to consider packets with a too large Forwarding Delay as being lost, as is required for some applications (e.g. voice, video, etc.). The Forwarding Delay Threshold is a parameter of the methodology, if it is applied it MUST be reported.

Measurement Units: seconds

Issues: None

See Also:

Convergence Packet Loss, Connectivity Packet Loss

3.8. Miscellaneous Terms

3.8.1. Stale Forwarding

Definition:

Forwarding of traffic to route entries that no longer exist or to route entries with next-hops that are no longer preferred.

Discussion:

Stale Forwarding can be caused by a Convergence Event and can manifest as a "black-hole" or microloop that produces packet loss, or out-of-order packets, or delayed packets. Stale Forwarding can exist until Network Convergence is completed.

Measurement Units: N/A

Issues: None

See Also: Network Convergence

3.8.2. Nested Convergence Event

Definition:

The occurrence of a Convergence Event while the route table is converging from a prior Convergence Event.

Discussion:

The Convergence Events for a Nested Convergence Event MUST occur with different neighbors. A possible observation from a Nested Convergence Event will be the withdrawal of routes from one neighbor while the routes of another neighbor are being installed.

Measurement Units: N/A

Issues: None

See Also: Convergence Event

4. 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.

5. IANA Considerations

This document requires no IANA considerations.

6. Acknowledgements

Thanks to Sue Hares, Al Morton, Kevin Dubray, Ron Bonica, David Ward, Peter De Vriendt, Anuj Dewagan and the BMWG for their contributions to this work.

7. References

7.1. Normative References

[Br91] Bradner, S., "Benchmarking terminology for network interconnection devices", RFC 1242, July 1991.

[Br97] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

[Br99] Bradner, S. and J. McQuaid, "Benchmarking Methodology for Network Interconnect Devices", RFC 2544, March 1999.

[Ca90] Callon, R., "Use of OSI IS-IS for routing in TCP/IP and dual environments", RFC 1195, December 1990.

[Co08] Coltun, R., Ferguson, D., Moy, J., and A. Lindem, "OSPF for IPv6", RFC 5340, July 2008.

[De02] Demichelis, C. and P. Chimento, "IP Packet Delay Variation Metric for IP Performance Metrics (IPPM)", RFC 3393, November 2002.

[Ho08] Hopps, C., "Routing IPv6 with IS-IS", RFC 5308, October 2008.

[Ko02] Koodli, R. and R. Ravikanth, "One-way Loss Pattern Sample Metrics", RFC 3357, August 2002.

[Ma98] Mandeville, R., "Benchmarking Terminology for LAN Switching Devices", RFC 2285, February 1998.

[Mo06] Morton, A., Ciavattone, L., Ramachandran, G., Shalunov, S., and J. Perser, "Packet Reordering Metrics", RFC 4737, November 2006.

[Mo98] Moy, J., "OSPF Version 2", STD 54, RFC 2328, April 1998.

[Po06] Poretsky, S., Perser, J., Erramilli, S., and S. Khurana, "Terminology for Benchmarking Network-layer Traffic Control Mechanisms", RFC 4689, October 2006.

[Po09a] Poretsky, S., "Considerations for Benchmarking Link-State IGP Data Plane Route Convergence", draft-ietf-bmwg-igp-dataplane-conv-app-17 (work in progress), March 2009.

[Po10m] Poretsky, S., Imhoff, B., and K. Michielsen, "Benchmarking Methodology for Link-State IGP Data Plane Route Convergence", draft-ietf-bmwg-igp-dataplane-conv-meth-20 (work in progress), March 2010.

7.2. Informative References

[Ca01] Casner, S., Alaettinoglu, C., and C. Kuan, "A Fine-Grained View of High Performance Networking", NANOG 22, June 2001.

[Ci03] Ciavattone, L., Morton, A., and G. Ramachandran, "Standardized Active Measurements on a Tier 1 IP Backbone", IEEE Communications Magazine p90-97, May 2003.

Authors' Addresses

Scott Poretsky
Allot Communications
67 South Bedford Street, Suite 400
Burlington, MA 01803
USA

Phone: + 1 508 309 2179
Email: sporetsky@allot.com

Brent Imhoff
Juniper Networks
1194 North Mathilda Ave
Sunnyvale, CA 94089
USA

Phone: + 1 314 378 2571
Email: bimhoff@planetspork.com

Kris Michielsen
Cisco Systems
6A De Kleetlaan
Diegem, BRABANT 1831
Belgium

Email: kmichiel@cisco.com

Network Working Group
Internet-Draft
Intended status: Informational
Expires: May 28, 2013

R. Papneja
Huawei Technologies
S. Vapiwala
J. Karthik
Cisco Systems
S. Poretsky
Allot Communications
S. Rao
Qwest Communications
JL. Le Roux
France Telecom

November 29, 2012

Methodology for Benchmarking MPLS-TE Fast Reroute Protection
draft-ietf-bmwg-protection-meth-14.txt

Abstract

This draft describes the methodology for benchmarking MPLS Fast Reroute (FRR) protection mechanisms for link and node protection. This document provides test methodologies and testbed setup for measuring failover times of Fast Reroute techniques while considering factors (such as underlying links) that might impact recovery times for real-time applications bound to MPLS traffic engineered (MPLS-TE) tunnels.

Status of this Memo

Copyright Notice

Table of Contents

1. Introduction

This document describes the methodology for benchmarking MPLS Fast Reroute (FRR) protection mechanisms. This document uses much of the terminology defined in [RFC 6414].

Protection mechanisms provide recovery of client services from a planned or an unplanned link or node failures. MPLS FRR protection mechanisms are generally deployed in a network infrastructure where MPLS is used for provisioning of point-to-point traffic engineered tunnels (tunnel). MPLS FRR protection mechanisms aim to reduce service disruption period by minimizing recovery time from most common failures.

Network elements from different manufacturers behave differently to network failures, which impacts the network’s ability and performance for failure recovery. It therefore becomes imperative for service providers to have a common benchmark to understand the performance behaviors of network elements.

There are two factors impacting service availability: frequency of failures and duration for which the failures persist. Failures can be classified further into two types: correlated and uncorrelated. Correlated and uncorrelated failures may be planned or unplanned.

Planned failures are generally predictable. Network implementations should be able to handle both planned and unplanned failures and recover gracefully within a time frame to maintain service assurance. Hence, failover recovery time is one of the most important benchmark that a service provider considers in choosing the building blocks for their network infrastructure.

A correlated failure is a result of the occurrence of two or more failures. A typical example is failure of a logical resource (e.g. layer-2 links) due to a dependency on a common physical resource (e.g. common conduit) that fails. Within the context of MPLS protection mechanisms, failures that arise due to Shared Risk Link Groups (SRLG) [RFC 4202] can be considered as correlated failures.

MPLS FRR [RFC 4090] allows for the possibility that the Label Switched Paths can be re-optimized in the minutes following Failover. IP Traffic would be re-routed according to the preferred path for the post-failure topology. Thus, MPLS-FRR may include additional steps following the occurrence of the failure detection [RFC 6414] and failover event [RFC 6414].

(1) Failover Event - Primary Path (Working Path) fails

(2) Failure Detection- Failover Event is detected

(3)

  a. Failover - Working Path switched to Backup path

  b. Re-Optimization of Working Path (possible change from Backup Path)

(4) Restoration [RFC 6414]

(5) Reversion [RFC 6414]

2. Document Scope

This document provides detailed test cases along with different topologies and scenarios that should be considered to effectively benchmark MPLS FRR protection mechanisms and failover times on the Data Plane. Different Failover Events and scaling considerations are also provided in this document.

All benchmarking test-cases defined in this document apply to Facility backup [RFC 4090]. The test cases cover set of interesting failure scenarios and the associated procedures benchmark the performance of the Device Under Test (DUT) to recover from failures. Data plane traffic is used to benchmark failover times. Testing scenarios related to MPLS-TE protection mechanisms when applied to MPLS Transport Profile and IP fast reroute applied to MPLS networks were not considered and are out of scope of this document. However, the test setups considered for MPLS based Layer 3 and Layer 2 services consider LDP over MPLS RSVP-TE configurations.

Benchmarking of correlated failures is out of scope of this document. Detection using Bi-directional Forwarding Detection (BFD) is outside the scope of this document, but mentioned in discussion sections.

The Performance of control plane is outside the scope of this benchmarking.

As described above, MPLS-FRR may include a Re-optimization of the Working Path, with possible packet transfer impairments. Characterization of Re-optimization is beyond the scope of this memo.

3. Existing Definitions and 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 BCP 14, [RFC 2119]. While [RFC 2119] defines the use of these key words primarily for Standards Track documents however, this Informational track document may use some of uses these keywords.

The reader is assumed to be familiar with the commonly used MPLS terminology, some of which is defined in [RFC 4090].

This document uses much of the terminology defined in [RFC 6414]. This document also uses existing terminology defined in other BMWG Work [RFC 1242], [RFC 2285], [RFC 4689]. Appendix B provide abbreviations used in the document

4. General Reference Topology

Figure 1 illustrates the basic reference testbed and is applicable to all the test cases defined in this document. The Tester is comprised of a Traffic Generator (TG) & Test Analyzer (TA) and Emulator. A Tester is connected to the test network and depending upon the test case, the DUT could vary. The Tester sends and receives IP traffic to the tunnel ingress and performs signaling protocol emulation to simulate real network scenarios in a lab environment. The Tester may also support MPLS-TE signaling to act as the ingress node to the MPLS tunnel. The lines in figures represent physical connections.

```
          +---------------------------+
          |              +------------|---------------+
          |              |            |               |
          |              |            |               |
      +--------+     +--------+    +--------+    +--------+   +--------+
  TG--|   R1   |-----|   R2   |----|   R3   |    |   R4   |   |  R5    |
      |        |-----|        |----|        |----|        |---|        |
      +--------+     +--------+    +--------+    +--------+   +--------+
          |             |              |             |            |
          |             |              |             |            |
          |         +--------+         |             |           TA
          +---------|   R6   |---------+             |
                    |        |----------------------+
                    +--------+
```

Fig. 1 Fast Reroute Topology

The tester MUST record the number of lost, duplicate, and out-of-order packets. It should further record arrival and departure times so that Failover Time, Additive Latency, and Reversion Time can be measured. The tester may be a single device or a test system emulating all the different roles along a primary or backup path.

The label stack is dependent of the following 3 entities:

(1) Type of protection (Link Vs Node)

(2) # of remaining hops of the primary tunnel from the PLR[RFC 6414]

(3) # of remaining hops of the backup tunnel from the PLR

Due to this dependency, it is RECOMMENDED that the benchmarking of failover times be performed on all the topologies provided in section 6.

5. Test Considerations

This section discusses the fundamentals of MPLS Protection testing:

(1) The types of network events that causes failover (section 5.1)

(2) Indications for failover (section 5.2)

(3) the use of data traffic (section 5.3)

(4) LSP Scaling (Section 5.4)

(5) IGP Selection (Section 5.5)

(6) Reversion of LSP (Section 5.6)

(7) Traffic generation (section 5.7)

5.1. Failover Events [RFC 6414]

The failover to the backup tunnel is primarily triggered by either link or node failures observed downstream of the Point of Local repair (PLR). The failure events are listed below.

Link Failure Events

- Interface Shutdown on PLR side with physical/link Alarm
- Interface Shutdown on remote side with physical/link Alarm
- Interface Shutdown on PLR side with RSVP hello enabled
- Interface Shutdown on remote side with RSVP hello enabled
- Interface Shutdown on PLR side with BFD
- Interface Shutdown on remote side with BFD
- Fiber Pull on the PLR side (Both TX & RX or just the TX)
- Fiber Pull on the remote side (Both TX & RX or just the RX)
- Online insertion and removal (OIR) on PLR side
- OIR on remote side
- Sub-interface failure on PLR side (e.g. shutting down of a VLAN)
- Sub-interface failure on remote side
- Parent interface shutdown on PLR side (an interface bearing multiple sub-interfaces)
- Parent interface shutdown on remote side

Node Failure Events

- A System reload initiated either by a graceful shutdown or by a power failure.
- A system crash due to a software failure or an assert.

5.2. Failure Detection [RFC 6414]

Link failure detection time depends on the link type and failure detection protocols running. For SONET/SDH, the alarm type (such as LOS, AIS, or RDI) can be used. Other link types have layer-two alarms, but they may not provide a short enough failure detection time. Ethernet based links enabled with MPLS/IP do not have layer 2 failure indicators, and therefore relies on layer 3 signaling for failure detection. However for directly connected devices, remote fault indication in the ethernet auto-negotiation scheme could be considered as a type of layer 2 link failure indicator.

MPLS has different failure detection techniques such as BFD, or use of RSVP hellos. These methods can be used for the layer 3 failure indicators required by Ethernet based links, or for some other non-Ethernet based links to help improve failure detection time. However, these fast failure detection mechanisms are out of scope.

The test procedures in this document can be used for a local failure or remote failure scenarios for comprehensive benchmarking and to evaluate failover performance independent of the failure detection techniques.

5.3. Use of Data Traffic for MPLS Protection benchmarking

Currently end customers use packet loss as a key metric for Failover Time [RFC 6414]. Failover Packet Loss [RFC 6414] is an externally observable event and has direct impact on application performance. MPLS protection is expected to minimize the packet loss in the event of a failure. For this reason it is important to develop a standard router benchmarking methodology for measuring MPLS protection that uses packet loss as a metric. At a known rate of forwarding, packet loss can be measured and the failover time can be determined. Measurement of control plane signaling to establish backup paths is not enough to verify failover. Failover is best determined when packets are actually traversing the backup path.

An additional benefit of using packet loss for calculation of failover time is that it allows use of a black-box test environment. Data traffic is offered at line-rate to the device under test (DUT) an emulated network failure event is forced to occur, and packet loss is externally measured to calculate the convergence time. This setup is independent of the DUT architecture.

In addition, this methodology considers the packets in error and duplicate packets [RFC 4689] that could have been generated during the failover process. The methodologies consider lost, out-of-order [RFC 4689] and duplicate packets to be impaired packets that contribute to the Failover Time.

5.4. LSP and Route Scaling

Failover time performance may vary with the number of established primary and backup tunnel label switched paths (LSP) and installed routes. However the procedure outlined here should be used for any number of LSPs (L) and number of routes protected by PLR(R). The amount of L and R must be recorded.

5.5. Selection of IGP

The underlying IGP could be ISIS-TE or OSPF-TE for the methodology proposed here. See [RFC 6412] for IGP options to consider and report.

5.6. Restoration and Reversion [RFC 6414]

Path restoration provides a method to restore an alternate primary LSP upon failure and to switch traffic from the Backup Path to the restored Primary Path (Reversion). In MPLS-FRR, Reversion can be implemented as Global Reversion or Local Reversion. It is important to include Restoration and Reversion as a step in each test case to measure the amount of packet loss, out of order packets, or duplicate packets that is produced.

Note: In addition to restoration and reversion, re-optimization can take place while the failure is still not recovered but it depends on the user configuration, and re-optimization timers.

5.7. Offered Load

It is suggested that there be three or more traffic streams as long as there is a steady and constant rate of flow for all the streams. In order to monitor the DUT performance for recovery times, a set of route prefixes should be advertised before traffic is sent. The traffic should be configured towards these routes.

Prefix-dependency behaviors are key in IP and tests with route-specific flows spread across the routing table will reveal this dependency. Generating traffic to all of the prefixes reachable by the protected tunnel (probably in a Round-Robin fashion, where the traffic is destined to all the prefixes but one prefix at a time in a cyclic manner) is not recommended. Round-Robin traffic generation is not recommended to all prefixes, as time to hit all the prefixes may be higher than the failover time. This phenomenon will reduce the granularity of the measured results and the results observed may not be accurate.

5.8. Tester Capabilities

It is RECOMMENDED that the Tester used to execute each test case have the following capabilities:

1.Ability to establish MPLS-TE tunnels and push/pop labels.

2.Ability to produce Failover Event [RFC 6414].

3.Ability to insert a timestamp in each data packet’s IP payload.

4.An internal time clock to control timestamping, time measurements, and time calculations.

5.Ability to disable or tune specific Layer-2 and Layer-3 protocol functions on any interface(s).

6.Ability to react upon the receipt of path error from the PLR

The Tester MAY be capable to make non-data plane convergence observations and use those observations for measurements.

5.9. Failover Time Measurement Methods

Failover Time is calculated using one of the following three methods

1. Packet-Loss Based method (PLBM): (Number of packets dropped/ packets per second * 1000) milliseconds. This method could also be referred as Loss-Derived method.

2. Time-Based Loss Method (TBLM): This method relies on the ability of the Traffic generators to provide statistics which reveal the duration of failure in milliseconds based on when the packet loss occurred (interval between non-zero packet loss and zero loss).

3. Timestamp Based Method (TBM): This method of failover calculation is based on the timestamp that gets transmitted as payload in the packets originated by the generator. The Traffic Analyzer records the timestamp of the last packet received before the failover event and the first packet after the failover and derives the time based on the difference between these 2 timestamps. Note: The payload could also contain sequence numbers for out-of-order packet calculation and duplicate packets.

The timestamp based method method would be able to detect Reversion impairments beyond loss, thus it is RECOMMENDED method as a Failover Time method.

6. Reference Test Setup

In addition to the general reference topology shown in figure 1, this section provides detailed insight into various proposed test setups that should be considered for comprehensively benchmarking the failover time in different roles along the primary tunnel

This section proposes a set of topologies that covers all the scenarios for local protection. All of these topologies can be mapped to the reference topology shown in Figure 1. Topologies provided in this section refer to the testbed required to benchmark failover time when the DUT is configured as a PLR in either Headend or midpoint role. Provided with each topology below is the label stack at the PLR. Penultimate Hop Popping (PHP) MAY be used and must be reported when used.

Figures 2 thru 9 use the following convention and are subset of figure 1:

a) HE is Headend
b) TE is Tail-End
c) MID is Mid point
d) MP is Merge Point
e) PLR is Point of Local Repair
f) PRI is Primary Path
g) BKP denotes Backup Path and Nodes
h) UR is Upstream Router

6.1. Link Protection

6.1.1. Link Protection - 1 hop primary (from PLR) and 1 hop backup TE tunnels

```
+-------+  +--------+    +--------+
|  R1   |  |   R2   | PRI|   R3   |
| UR/HE |--| HE/MID |----|  MP/TE |
|       |  |  PLR   |----|        |
+-------+  +--------+ BKP+--------+
```

Figure 2.

Traffic	Num of Labels before failure	Num of labels after failure
IP TRAFFIC (P-P)	0	0
Layer3 VPN (PE-PE)	1	1
Layer3 VPN (PE-P)	2	2
Layer2 VC (PE-PE)	1	1
Layer2 VC (PE-P)	2	2
Mid-point LSPs	0	0

Note: Please note the following:

a) For P-P case, R2 and R3 acts as P routers
b) For PE-PE case,R2 acts as PE and R3 acts as a remote PE
c) For PE-P case,R2 acts as a PE router, R3 acts as a P router and R5 acts as remote PE router (Please refer to figure 1 for complete setup)
d) For Mid-point case, R1, R2 and R3 act as shown in above figure HE, Midpoint/PLR and TE respectively

6.1.2. Link Protection - 1 hop primary (from PLR) and 2 hop backup TE tunnels

```
+-------+    +--------+    +--------+
|  R1   |    |  R2    |    |   R3   |
| UR/HE |    | HE/MID |PRI |  MP/TE |
|       |----|  PLR   |----|        |
+-------+    +--------+    +--------+
                 |BKP                |
                 |    +--------+     |
                 |    |   R6   |     |
                 |----|  BKP   |-----|
                      |   MID  |
                      +--------+
```

Figure 3.

Traffic	Num of Labels before failure	Num of labels after failure
IP TRAFFIC (P-P)	0	1
Layer3 VPN (PE-PE)	1	2
Layer3 VPN (PE-P)	2	3
Layer2 VC (PE-PE)	1	2
Layer2 VC (PE-P)	2	3
Mid-point LSPs	0	1

Note: Please note the following:

a) For P-P case, R2 and R3 acts as P routers
b) For PE-PE case,R2 acts as PE and R3 acts as a remote PE
c) For PE-P case,R2 acts as a PE router, R3 acts as a P router and R5 acts as remote PE router (Please refer to figure 1 for complete setup)
d) For Mid-point case, R1, R2 and R3 act as shown in above figure HE, Midpoint/PLR and TE respectively

6.1.3. Link Protection - 2+ hop (from PLR) primary and 1 hop backup TE tunnels

```
+--------+    +--------+    +--------+      +--------+
|  R1    |    | R2     |PRI |   R3   |PRI   |   R4   |
|  UR/HE |----| HE/MID |----| MP/MID |------|   TE   |
|        |    | PLR    |----|        |      |        |
+--------+    +--------+ BKP+--------+      +--------+
```

Figure 4.

Traffic	Num of Labels before failure	Num of labels after failure
IP TRAFFIC (P-P)	1	1
Layer3 VPN (PE-PE)	2	2
Layer3 VPN (PE-P)	3	3
Layer2 VC (PE-PE)	2	2
Layer2 VC (PE-P)	3	3
Mid-point LSPs	1	1

Note: Please note the following:

a) For P-P case, R2, R3 and R4 acts as P routers
b) For PE-PE case,R2 acts as PE and R4 acts as a remote PE
c) For PE-P case,R2 acts as a PE router, R3 acts as a P router and R5 acts as remote PE router (Please refer to figure 1 for complete setup)
d) For Mid-point case, R1, R2, R3 and R4 act as shown in above figure HE, Midpoint/PLR and TE respectively

6.1.4. Link Protection - 2+ hop (from PLR) primary and 2 hop backup TE tunnels

```
+--------+    +--------+PRI +--------+  PRI +--------+
|  R1    |    |  R2    |    |   R3   |      |   R4   |
| UR/HE  |----| HE/MID |----|  MP/MID|------|   TE   |
|        |    | PLR    |    |        |      |        |
+--------+    +--------+    +--------+      +--------+
              BKP|              |
                 |   +--------+ |
                 |   |   R6   | |
                 +---|  BKP   |-
                     |  MID   |
                     +--------+
```

Figure 5.

Traffic	Num of Labels before failure	Num of labels after failure
IP TRAFFIC (P-P)	1	2
Layer3 VPN (PE-PE)	2	3
Layer3 VPN (PE-P)	3	4
Layer2 VC (PE-PE)	2	3
Layer2 VC (PE-P)	3	4
Mid-point LSPs	1	2

Note: Please note the following:

a) For P-P case, R2, R3 and R4 acts as P routers
b) For PE-PE case,R2 acts as PE and R4 acts as a remote PE
c) For PE-P case,R2 acts as a PE router, R3 acts as a P router and R5 acts as remote PE router (Please refer to figure 1 for complete setup)
d) For Mid-point case, R1, R2, R3 and R4 act as shown in above figure HE, Midpoint/PLR and TE respectively

6.2. Node Protection

6.2.1. Node Protection - 2 hop primary (from PLR) and 1 hop backup TE tunnels

```
+--------+    +--------+    +--------+      +--------+
|  R1    |    |  R2    |PRI |   R3   | PRI  |   R4   |
| UR/HE  |----| HE/MID |----|  MID   |------|  MP/TE |
|        |    |  PLR   |    |        |      |        |
+--------+    +--------+    +--------+      +--------+
                |BKP                          |
                 -----------------------------
```

Figure 6.

Traffic	Num of Labels before failure	Num of labels after failure
IP TRAFFIC (P-P)	1	0
Layer3 VPN (PE-PE)	2	1
Layer3 VPN (PE-P)	3	2
Layer2 VC (PE-PE)	2	1
Layer2 VC (PE-P)	3	2
Mid-point LSPs	1	0

Note: Please note the following:

a) For P-P case, R2, R3 and R3 acts as P routers
b) For PE-PE case,R2 acts as PE and R4 acts as a remote PE
c) For PE-P case,R2 acts as a PE router, R4 acts as a P router and R5 acts as remote PE router (Please refer to figure 1 for complete setup)
d) For Mid-point case, R1, R2, R3 and R4 act as shown in above figure HE, Midpoint/PLR and TE respectively

6.2.2. Node Protection - 2 hop primary (from PLR) and 2 hop backup TE tunnels

```
+--------+    +--------+    +--------+    +--------+
|  R1    |    |  R2    |    |   R3   |    |   R4   |
| UR/HE  |    | HE/MID |PRI |  MID   |PRI |  MP/TE |
|        |----|  PLR   |----|        |----|        |
+--------+    +--------+    +--------+    +--------+
                |                            |
             BKP|         +--------+         |
                |         |   R6   |         |
                 ---------|  BKP   |---------
                          |  MID   |
                          +--------+
```

Figure 7.

Traffic	Num of Labels before failure	Num of labels after failure
IP TRAFFIC (P-P)	1	1
Layer3 VPN (PE-PE)	2	2
Layer3 VPN (PE-P)	3	3
Layer2 VC (PE-PE)	2	2
Layer2 VC (PE-P)	3	3
Mid-point LSPs	1	1

Note: Please note the following:

a) For P-P case, R2, R3 and R4 acts as P routers
b) For PE-PE case,R2 acts as PE and R4 acts as a remote PE
c) For PE-P case,R2 acts as a PE router, R4 acts as a P router and R5 ac
ts as remote
PE router (Please refer to figure 1 for complete setup)
d) For Mid-point case, R1, R2, R3 and R4 act as shown in above figure HE
, Midpoint/PLR
and TE respectively

6.2.3. Node Protection - 3+ hop primary (from PLR) and 1 hop backup TE tunnels

```
+--------+  +--------+PRI+--------+PRI+--------+PRI+--------+
|  R1    |  |  R2    |   |   R3   |   |   R4   |   |   R5   |
| UR/HE  |--| HE/MID |---| MID    |---|  MP    |---|  TE    |
|        |  |  PLR   |   |        |   |        |   |        |
+--------+  +--------+   +--------+   +--------+   +--------+
            BKP|                          |
                --------------------------
```

Figure 8.

Traffic	Num of Labels before failure	Num of labels after failure
IP TRAFFIC (P-P)	1	1
Layer3 VPN (PE-PE)	2	2
Layer3 VPN (PE-P)	3	3
Layer2 VC (PE-PE)	2	2
Layer2 VC (PE-P)	3	3
Mid-point LSPs	1	1

Note: Please note the following:

a) For P-P case, R2, R3, R4 and R5 acts as P routers
b) For PE-PE case,R2 acts as PE and R5 acts as a remote PE
c) For PE-P case,R2 acts as a PE router, R4 acts as a P router and R5 acts as remote
PE router (Please refer to figure 1 for complete setup)
d) For Mid-point case, R1, R2, R3, R4 and R5 act as shown in above figure HE,
Midpoint/PLR and TE respectively

6.2.4. Node Protection - 3+ hop primary (from PLR) and 2 hop backup TE tunnels

```
+--------+   +--------+   +--------+   +--------+   +--------+
|  R1    |   |  R2    |   |   R3   |   |   R4   |   |   R5   |
| UR/HE  |   | HE/MID |PRI|  MID   |PRI|  MP    |PRI|  TE    |
|        |-- |  PLR   |---|        |---|        |---|        |
+--------+   +--------+   +--------+   +--------+   +--------+
              BKP|                             |
                 |         +--------+          |
                 |         |  R6    |          |
                  ---------|  BKP   |-------
                           |  MID   |
                           +--------+
```

Figure 9.

Traffic	Num of Labels before failure	Num of labels after failure
IP TRAFFIC (P-P)	1	2
Layer3 VPN (PE-PE)	2	3
Layer3 VPN (PE-P)	3	4
Layer2 VC (PE-PE)	2	3
Layer2 VC (PE-P)	3	4
Mid-point LSPs	1	2

Note: Please note the following:

a) For P-P case, R2, R3, R4 and R5 acts as P routers
b) For PE-PE case,R2 acts as PE and R5 acts as a remote PE
c) For PE-P case,R2 acts as a PE router, R4 acts as a P router and R5 acts as remote
PE router (Please refer to figure 1 for complete setup)
d) For Mid-point case, R1, R2, R3, R4 and R5 act as shown in above figure HE,
Midpoint/PLR and TE respectively

7. Test Methodology

The procedure described in this section can be applied to all the 8 base test cases and the associated topologies. The backup as well as the primary tunnels are configured to be alike in terms of bandwidth usage. In order to benchmark failover with all possible label stack depth applicable as seen with current deployments, it is RECOMMENDED to perform all of the test cases provided in this section. The forwarding performance test cases in section 7.1 MUST be performed prior to performing the failover test cases.

The considerations of Section 4 of [RFC 2544] are applicable when evaluating the results obtained using these methodologies as well.

7.1. MPLS FRR Forwarding Performance

Benchmarking Failover Time [RFC 6414] for MPLS protection first requires baseline measurement of the forwarding performance of the test topology including the DUT. Forwarding performance is benchmarked by the Throughput as defined in [RFC 5695] and measured in units pps. This section provides two test cases to benchmark forwarding performance. These are with the DUT configured as a Headend PLR, Mid-Point PLR, and Egress PLR.

7.1.1. Headend PLR Forwarding Performance

Objective:

To benchmark the maximum rate (pps) on the PLR (as headend) over primary LSP and backup LSP.

Test Setup:

A. Select any one topology out of the 8 from section 6.

B. Select or enable IP, Layer 3 VPN or Layer 2 VPN services with DUT as Headend PLR.

C. The DUT will also have 2 interfaces connected to the traffic Generator/analyzer. (If the node downstream of the PLR is not a simulated node, then the Ingress of the tunnel should have one link connected to the traffic generator and the node downstream to the PLR or the egress of the tunnel should have a link connected to the traffic analyzer).

Procedure:

1. Establish the primary LSP on R2 required by the topology selected.

2. Establish the backup LSP on R2 required by the selected topology.

3. Verify primary and backup LSPs are up and that primary is protected.

4. Verify Fast Reroute protection is enabled and ready.

5. Setup traffic streams as described in section 5.7.

6. Send MPLS traffic over the primary LSP at the Throughput supported by the DUT (section 6, RFC 2544).

7. Record the Throughput over the primary LSP.

8. Trigger a link failure as described in section 5.1.

9. Verify that the offered load gets mapped to the backup tunnel and measure the Additive Backup Delay (RFC 6414).

10. 30 seconds after Failover, stop the offered load and measure the Throughput, Packet Loss, Out-of-Order Packets, and Duplicate Packets over the Backup LSP.

11. Adjust the offered load and repeat steps 6 through 10 until the Throughput values for the primary and backup LSPs are equal.

12. Record the final Throughput, which corresponds to the offered load that will be used for the Headend PLR failover test cases.

7.1.2. Mid-Point PLR Forwarding Performance

Objective:

To benchmark the maximum rate (pps) on the PLR (as mid-point) over primary LSP and backup LSP.

Test Setup:

A. Select any one topology out of the 8 from section 6.

B. The DUT will also have 2 interfaces connected to the traffic generator.

Procedure:

1. Establish the primary LSP on R1 required by the topology selected.

2. Establish the backup LSP on R2 required by the selected topology.

3. Verify primary and backup LSPs are up and that primary is protected.

4. Verify Fast Reroute protection is enabled and ready.

5. Setup traffic streams as described in section 5.7.

6. Send MPLS traffic over the primary LSP at the Throughput supported by the DUT (section 6, RFC 2544).

7. Record the Throughput over the primary LSP.

8. Trigger a link failure as described in section 5.1.

9. Verify that the offered load gets mapped to the backup tunnel and measure the Additive Backup Delay (RFC 6414).

10. 30 seconds after Failover, stop the offered load and measure the Throughput, Packet Loss, Out-of-Order Packets, and Duplicate Packets over the Backup LSP.

11. Adjust the offered load and repeat steps 6 through 10 until the Throughput values for the primary and backup LSPs are equal.

12. Record the final Throughput which corresponds to the offered load that will be used for the Mid-Point PLR failover test cases.

7.2. Headend PLR with Link Failure

Objective:

To benchmark the MPLS failover time due to link failure events described in section 5.1 experienced by the DUT which is the Headend PLR.

Test Setup:

A. Select any one topology out of the 8 from section 6.

B. Select or enable IP, Layer 3 VPN or Layer 2 VPN services with DUT as Headend PLR.

C. The DUT will also have 2 interfaces connected to the traffic Generator/analyzer. (If the node downstream of the PLR is not a simulated node, then the Ingress of the tunnel should have one link connected to the traffic generator and the node downstream to the PLR or the egress of the tunnel should have a link connected to the traffic analyzer).

Test Configuration:

1. Configure the number of primaries on R2 and the backups on R2 as required by the topology selected.

2. Configure the test setup to support Reversion.

3. Advertise prefixes (as per FRR Scalability Table described in Appendix A) by the tail end.

Procedure:

Test Case "7.1.1. Headend PLR Forwarding Performance" MUST be completed first to obtain the Throughput to use as the offered load.

1. Establish the primary LSP on R2 required by the topology selected.

2. Establish the backup LSP on R2 required by the selected topology.

3. Verify primary and backup LSPs are up and that primary is protected.

4. Verify Fast Reroute protection is enabled and ready.

5. Setup traffic streams for the offered load as described in section 5.7.

6. Provide the offered load from the tester at the Throughput [RFC 1242] level obtained from test case 7.1.1.

7. Verify traffic is switched over Primary LSP without packet loss.

8. Trigger a link failure as described in section 5.1.

9. Verify that the offered load gets mapped to the backup tunnel and measure the Additive Backup Delay.

10. 30 seconds after Failover [RFC 6414], stop the offered load and measure the total Failover Packet Loss [RFC 6414].

11. Calculate the Failover Time [RFC 6414] benchmark using the selected Failover Time Calculation Method (TBLM, PLBM, or TBM) [RFC 6414].

12. Restart the offered load and restore the primary LSP to verify Reversion [RFC 6414] occurs and measure the Reversion Packet Loss [RFC 6414].

13. Calculate the Reversion Time [RFC 6414] benchmark using the selected Failover Time Calculation Method (TBLM, PLBM, or TBM) [RFC 6414].

14. Verify Headend signals new LSP and protection should be in place again.

IT is RECOMMENDED that this procedure be repeated for each of the link failure triggers defined in section 5.1.

7.3. Mid-Point PLR with Link Failure

Objective:

To benchmark the MPLS failover time due to link failure events described in section 5.1 experienced by the DUT which is the Mid-Point PLR.

Test Setup:

A. Select any one topology out of the 8 from section 6.

B. The DUT will also have 2 interfaces connected to the traffic generator.

Test Configuration:

1. Configure the number of primaries on R1 and the backups on R2 as required by the topology selected.

2. Configure the test setup to support Reversion.

3. Advertise prefixes (as per FRR Scalability Table described in Appendix A) by the tail end.

Procedure:

Test Case "7.1.2. Mid-Point PLR Forwarding Performance" MUST be completed first to obtain the Throughput to use as the offered load.

1. Establish the primary LSP on R1 required by the topology selected.

2. Establish the backup LSP on R2 required by the selected topology.

3. Perform steps 3 through 14 from section 7.2 Headend PLR with Link Failure.

IT is RECOMMENDED that this procedure be repeated for each of the link failure triggers defined in section 5.1.

7.4. Headend PLR with Node Failure

Objective:

To benchmark the MPLS failover time due to Node failure events described in section 5.1 experienced by the DUT which is the Headend PLR.

Test Setup:

A. Select any one topology out of the 8 from section 6.

B. Select or enable IP, Layer 3 VPN or Layer 2 VPN services with DUT as Headend PLR.

C. The DUT will also have 2 interfaces connected to the traffic generator/analyzer.

Test Configuration:

1. Configure the number of primaries on R2 and the backups on R2 as required by the topology selected.

2. Configure the test setup to support Reversion.

3. Advertise prefixes (as per FRR Scalability Table described in Appendix A) by the tail end.

Procedure:

Test Case "7.1.1. Headend PLR Forwarding Performance" MUST be completed first to obtain the Throughput to use as the offered load.

1. Establish the primary LSP on R2 required by the topology selected.

2. Establish the backup LSP on R2 required by the selected topology.

3. Verify primary and backup LSPs are up and that primary is protected.

4. Verify Fast Reroute protection is enabled and ready.

5. Setup traffic streams for the offered load as described in section 5.7.

6. Provide the offered load from the tester at the Throughput [RFC 1242] level obtained from test case 7.1.1.

7. Verify traffic is switched over Primary LSP without packet loss.

8. Trigger a node failure as described in section 5.1.

9. Perform steps 9 through 14 in 7.2 Headend PLR with Link Failure.

IT is RECOMMENDED that this procedure be repeated for each of the node failure triggers defined in section 5.1.

7.5. Mid-Point PLR with Node Failure

Objective:

To benchmark the MPLS failover time due to Node failure events described in section 5.1 experienced by the DUT which is the Mid-Point PLR.

Test Setup:

A. Select any one topology from section 6.1 to 6.2.

B. The DUT will also have 2 interfaces connected to the traffic generator.

Test Configuration:

1. Configure the number of primaries on R1 and the backups on R2 as required by the topology selected.

2. Configure the test setup to support Reversion.

3. Advertise prefixes (as per FRR Scalability Table described in Appendix A) by the tail end.

Procedure:

Test Case "7.1.1. Mid-Point PLR Forwarding Performance" MUST be completed first to obtain the Throughput to use as the offered load.

1. Establish the primary LSP on R1 required by the topology selected.

2. Establish the backup LSP on R2 required by the selected topology.

3. Verify primary and backup LSPs are up and that primary is protected.

4. Verify Fast Reroute protection is enabled and ready.

5. Setup traffic streams for the offered load as described in section 5.7.

6. Provide the offered load from the tester at the Throughput [RFC 1242] level obtained from test case 7.1.1.

7. Verify traffic is switched over Primary LSP without packet loss.

8. Trigger a node failure as described in section 5.1.

9. Perform steps 9 through 14 in 7.2 Headend PLR with Link Failure.

IT is RECOMMENDED that this procedure be repeated for each of the node failure triggers defined in section 5.1.

8. Reporting Format

For each test, it is RECOMMENDED that the results be reported in the following format.

Parameter	Units
IGP used for the test	ISIS-TE/ OSPF-TE

Interface types	Gige,POS,ATM,VLAN etc.
Packet Sizes offered to the DUT	Bytes (at layer 3)
Offered Load (Throughput)	packets per second
IGP routes advertised	Number of IGP routes
Penultimate Hop Popping	Used/Not Used
RSVP hello timers	Milliseconds
Number of Protected tunnels	Number of tunnels
Number of VPN routes installed on the Headend	Number of VPN routes
Number of VC tunnels	Number of VC tunnels
Number of mid-point tunnels	Number of tunnels
Number of Prefixes protected by Primary	Number of LSPs
Topology being used	Section number, and figure reference
Failover Event	Event type
Re-optimization	Yes/No

Benchmarks (to be recorded for each test case):

Failover-

Failover Time	seconds
Failover Packet Loss	packets
Additive Backup Delay	seconds
Out-of-Order Packets	packets
Duplicate Packets	packets
Failover Time Calculation Method	Method Used

Reversion-

Reversion Time	seconds
Reversion Packet Loss	packets
Additive Backup Delay	seconds
Out-of-Order Packets	packets
Duplicate Packets	packets
Failover Time Calculation Method	Method Used

9. 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.

10. IANA Considerations

This draft does not require any new allocations by IANA.

11. Acknowledgements

We would like to thank Jean Philip Vasseur for his invaluable input to the document, Curtis Villamizar for his contribution in suggesting text on definition and need for benchmarking Correlated failures and Bhavani Parise for his textual input and review. Additionally we would like to thank Al Morton, Arun Gandhi, Amrit Hanspal, Karu Ratnam, Raveesh Janardan, Andrey Kiselev, and Mohan Nanduri for their formal reviews of this document.

12. References

12.1. Informative References

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

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

[RFC 4202] Kompella, K., Rekhter, Y., "Routing Extensions in Support of Generalized Multi-Protocol Label Switching (GMPLS)", RFC 4202, October 2005.

12.2. Normative References

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

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

[RFC 4090] Pan, P., Swallow, G., and A. Atlas, "Fast Reroute Extensions to RSVP-TE for LSP Tunnels", RFC 4090, May 2005.

[RFC 5695] Akhter, A., Asati, R., and C. Pignataro, "MPLS Forwarding Benchmarking Methodology for IP Flows", RFC 5695, November 2009.

[RFC 6414] Poretsky, S., Papneja, R., Karthik, J., and S. Vapiwala, "Benchmarking Terminology for Protection Performance", RFC 6414, November 2011.

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

[RFC 6412] Poretsky, S., Imhoff, B., and K. Michielsen, "Terminology for Benchmarking Link-State IGP Data-Plane Route Convergence", RFC 6412, November 2011.

Appendix A. Fast Reroute Scalability Table

This section provides the recommended numbers for evaluating the scalability of fast reroute implementations. It also recommends the typical numbers for IGP/VPNv4 Prefixes, LSP Tunnels and VC entries. Based on the features supported by the device under test (DUT), appropriate scaling limits can be used for the test bed.

A1. FRR IGP Table

No. of Headend TE Tunnels	IGP Prefixes
1	100
1	500
1	1000
1	2000
1	5000
2 (Load Balance)	100
2 (Load Balance)	500
2 (Load Balance)	1000
2 (Load Balance)	2000
2 (Load Balance)	5000
100	100
500	500
1000	1000
2000	2000

A2. FRR VPN Table

No. of Headend TE Tunnels	VPNv4 Prefixes
1	100
1	500
1	1000
1	2000
1	5000
1	10000
1	20000
1	Max
2 (Load Balance)	100
2 (Load Balance)	500
2 (Load Balance)	1000
2 (Load Balance)	2000
2 (Load Balance)	5000
2 (Load Balance)	10000
2 (Load Balance)	20000
2 (Load Balance)	Max

A3. FRR Mid-Point LSP Table

No of Mid-point TE LSPs could be configured at recommended levels - 100, 500, 1000, 2000, or max supported number.

A2. FRR VC Table

No. of Headend TE Tunnels	VC entries
1	100
1	500
1	1000
1	2000
1	Max
100	100
500	500
1000	1000
2000	2000

Appendix B. Abbreviations

AIS - Alarm Indication Signal
BFD - Bidirectional Fault Detection
BGP - Border Gateway protocol
CE - Customer Edge
DUT - Device Under Test
FRR - Fast Reroute
IGP - Interior Gateway Protocol
IP - Internet Protocol
LOS - Loss of Signal
LSP - Label Switched Path
MP - Merge Point
MPLS - Multi Protocol Label Switching
N-Nhop - Next - Next Hop
Nhop - Next Hop
OIR - Online Insertion and Removal
P - Provider
PE - Provider Edge
PHP - Penultimate Hop Popping
PLR - Point of Local Repair
RSVP - Resource reSerVation Protocol
SRLG - Shared Risk Link Group
TA - Traffic Analyzer
TE - Traffic Engineering
TG - Traffic Generator
VC - Virtual Circuit
VPN - Virtual Private Network

Authors' Addresses

Rajiv Papneja
Huawei Technologies
2330 Central Expressway
Santa Clara, CA 95050
USA

Email: rajiv.papneja@huawei.com

Samir Vapiwala
Cisco Systems
300 Beaver Brook Road
Boxborough, MA 01719
USA

Email: svapiwal@cisco.com

Jay Karthik
Cisco Systems
300 Beaver Brook Road
Boxborough, MA 01719
USA

Email: jkarthik@cisco.com

Scott Poretsky
Allot Communications
USA

Email: sporetsky@allot.com

Shankar Rao
Qwest Communications
950 17th Street
Suite 1900
Denver, CO 80210
USA

Email: shankar.rao@du.edu

JL. Le Roux
France Telecom
2 av Pierre Marzin
22300 Lannion
France

Email: jeanlouis.leroux@orange.com

Benchmarking Methodology Working Group C. Davids
Internet-Draft Illinois Institute of Technology
Intended status: Informational V. Gurbani
Expires: May 16, 2015 Bell Laboratories,
Alcatel-Lucent
S. Poretsky
Allot Communications
November 12, 2014

Methodology for Benchmarking Session Initiation Protocol (SIP) Devices: Basic session setup and registration
draft-ietf-bmwg-sip-bench-meth-12

Abstract

This document provides a methodology for benchmarking the Session Initiation Protocol (SIP) performance of devices. Terminology related to benchmarking SIP devices is described in the companion terminology document. Using these two documents, benchmarks can be obtained and compared for different types of devices such as SIP Proxy Servers, Registrars and Session Border Controllers. The term "performance" in this context means the capacity of the device-under-test (DUT) to process SIP messages. Media streams are used only to study how they impact the signaling behavior. The intent of the two documents is to provide a normalized set of tests that will enable an objective comparison of the capacity of SIP devices. Test setup parameters and a methodology are necessary because SIP allows a wide range of configuration and operational conditions that can influence performance benchmark measurements.

Table of Contents

1. Terminology

In this document, the key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "NOT RECOMMENDED", "MAY", and "OPTIONAL" are to be interpreted as described in BCP 14, conforming to [RFC2119] and indicate requirement levels for compliant implementations.

RFC 2119 defines the use of these key words to help make the intent of standards track documents as clear as possible. While this document uses these keywords, this document is not a standards track document. The term Throughput is defined in [RFC2544].

Terms specific to SIP [RFC3261] performance benchmarking are defined in [I-D.sip-bench-term].

2. Introduction

This document describes the methodology for benchmarking Session Initiation Protocol (SIP) performance as described in the Terminology document [I-D.sip-bench-term]. The methodology and terminology are to be used for benchmarking signaling plane performance with varying signaling and media load. Media streams, when used, are used only to study how they impact the signaling behavior. This document concentrates on benchmarking SIP session setup and SIP registrations only.

The device-under-test (DUT) is a RFC3261-capable [RFC3261] network intermediary that plays the role of a registrar, redirect server, stateful proxy, a Session Border Controller (SBC) or a B2BUA. This document does not require the intermediary to assume the role of a stateless proxy. Benchmarks can be obtained and compared for different types of devices such as a SIP proxy server, Session Border Controllers (SBC), SIP registrars and a SIP proxy server paired with a media relay.

The test cases provide metrics for benchmarking the maximum 'SIP Registration Rate' and maximum 'SIP Session Establishment Rate' that the DUT can sustain over an extended period of time without failures (extended period of time is defined in the algorithm in Section 4.10). Some cases are included to cover encrypted SIP. The test topologies that can be used are described in the Test Setup section. Topologies in which the DUT handles media as well as those in which the DUT does not handle media are both considered. The measurement of the performance characteristics of the media itself is outside the scope of these documents.

Benchmark metrics could possibly be impacted by Associated Media. The selected values for Session Duration and Media Streams per Session enable benchmark metrics to be benchmarked without Associated Media. Session Setup Rate could possibly be impacted by the selected value for Maximum Sessions Attempted. The benchmark for Session Establishment Rate is measured with a fixed value for maximum Session Attempts.

Finally, the overall value of these tests is to serve as a comparison function between multiple SIP implementations. One way to use these tests is to derive benchmarks with SIP devices from Vendor-A, derive a new set of benchmarks with similar SIP devices from Vendor-B and perform a comparison on the results of Vendor-A and Vendor-B. This document does not make any claims on the interpretation of such results.

3. Benchmarking Topologies

Test organizations need to be aware that these tests generate large volumes of data and consequently ensure that networking devices like hubs, switches or routers are able to handle the generated volume.

The test cases enumerated in Section 6.1 to Section 6.6 operate on two test topologies: one in which the DUT does not process the media (Figure 1) and the other in which it does process media (Figure 2). In both cases, the tester or emulated agent (EA) sends traffic into the DUT and absorbs traffic from the DUT. The diagrams in Figure 1 and Figure 2 represent the logical flow of information and do not dictate a particular physical arrangement of the entities.

Figure 1 depicts a layout in which the DUT is an intermediary between the two interfaces of the EA. If the test case requires the exchange of media, the media does not flow through the DUT but rather passes directly between the two endpoints. Figure 2 shows the DUT as an intermediary between the two interfaces of the EA. If the test case requires the exchange of media, the media flows through the DUT between the endpoints.

```
      +--------+   Session   +--------+  Session    +--------+
      |        |   Attempt   |        |  Attempt    |        |
      |        |------------>+        |------------>+        |
      |        |             |        |             |        |
      |        |   Response  |        |  Response   |        |
      | Tester +<------------|  DUT   +<------------| Tester |
      |  (EA)  |             |        |             |  (EA)  |
      |        |             |        |             |        |
      +--------+             +--------+             +--------+
         /|\                                           /|\
          |              Media (optional)               |
          +==============================================+
```

Figure 1: DUT as an intermediary, end-to-end media

```
      +--------+   Session   +--------+  Session    +--------+
      |        |   Attempt   |        |  Attempt    |        |
      |        |------------>+        |------------>+        |
      |        |             |        |             |        |
      |        |   Response  |        |  Response   |        |
      | Tester +<------------|  DUT   +<------------| Tester |
      |  (EA)  |             |        |             |  (EA)  |
      |        |<===========>|        |<===========>|        |
      +--------+   Media     +--------+    Media    +--------+
                 (Optional)             (Optional)
```

Figure 2: DUT as an intermediary forwarding media

The test cases enumerated in Section 6.7 and Section 6.8 use the topology in Figure 3 below.

```
      +--------+ Registration +--------+
      |        |   request    |        |
      |        |------------->+        |
      |        |              |        |
      |        |   Response   |        |
      | Tester +<-------------|  DUT   |
      |  (EA)  |              |        |
      |        |              |        |
      +--------+              +--------+
```

Figure 3: Registration and Re-registration tests

During registration or re-registration, the DUT may involve backend network elements and data stores. These network elements and data stores are not shown in Figure 3, but it is understood that they will impact the time required for the DUT to generate a response.

This document explicitly separates a registration test (Section 6.7) from a re-registration test (Section 6.8) because in certain networks, the time to re-register may vary from the time to perform an initial registration due to the backend processing involved. It is expected that the registration tests and the re-registration test will be performed with the same set of backend network elements in order to derive a stable metric.

4. Test Setup Parameters

4.1. Selection of SIP Transport Protocol

Test cases may be performed with any transport protocol supported by SIP. This includes, but is not limited to, TCP, UDP, TLS and websockets. The protocol used for the SIP transport protocol must be reported with benchmarking results.

SIP allows a DUT to use different transports for signaling on either side of the connection to the EAs. Therefore, this document assumes that the same transport is used on both sides of the connection; if this is not the case in any of the tests, the transport on each side of the connection MUST be reported in the test reporting template.

4.2. Connection-oriented Transport Management

SIP allows a device to open one connection and send multiple requests over the same connection (responses are normally received over the same connection that the request was sent out on). The protocol also allows a device to open a new connection for each individual request. A connection management strategy will have an impact on the results obtained from the test cases, especially for connection-oriented transports such as TLS. For such transports, the cryptographic handshake must occur every time a connection is opened.

The connection management strategy, i.e., use of one connection to send all requests or closing an existing connection and opening a new connection to send each request, MUST be reported with the benchmarking result.

4.3. Signaling Server

The Signaling Server is defined in the companion terminology document, ([I-D.sip-bench-term], Section 3.2.2). The Signaling Server is a DUT.

4.4. Associated Media

Some tests require Associated Media to be present for each SIP session. The test topologies to be used when benchmarking DUT performance for Associated Media are shown in Figure 1 and Figure 2.

4.5. Selection of Associated Media Protocol

The test cases specified in this document provide SIP performance independent of the protocol used for the media stream. Any media protocol supported by SIP may be used. This includes, but is not limited to, RTP, and SRTP. The protocol used for Associated Media MUST be reported with benchmarking results.

4.6. Number of Associated Media Streams per SIP Session

Benchmarking results may vary with the number of media streams per SIP session. When benchmarking a DUT for voice, a single media stream is used. When benchmarking a DUT for voice and video, two media streams are used. The number of Associated Media Streams MUST be reported with benchmarking results.

4.7. Codec Type

The test cases specified in this document provide SIP performance independent of the media stream codec. Any codec supported by the EAs may be used. The codec used for Associated Media MUST be reported with the benchmarking results.

4.8. Session Duration

The value of the DUT’s performance benchmarks may vary with the duration of SIP sessions. Session Duration MUST be reported with benchmarking results. A Session Duration of zero seconds indicates transmission of a BYE immediately following a successful SIP establishment. Setting this parameter to the value ’0’ indicates that a BYE will be sent by the EA immediately after the EA receives a 200 OK to the INVITE. Setting this parameter to a time value greater than the duration of the test indicates that a BYE is never sent.

4.9. Attempted Sessions per Second (sps)

The value of the DUT's performance benchmarks may vary with the Session Attempt Rate offered by the tester. Session Attempt Rate MUST be reported with the benchmarking results.

The test cases enumerated in Section 6.1 to Section 6.6 require that the EA is configured to send the final 2xx-class response as quickly as it can. This document does not require the tester to add any delay between receiving a request and generating a final response.

4.10. Benchmarking algorithm

In order to benchmark the test cases uniformly in Section 6, the algorithm described in this section should be used. A prosaic description of the algorithm and a pseudo-code description are provided below, and a simulation written in the R statistical language [Rtool] is provided in Appendix A.

The goal is to find the largest value, R, a SIP Session Attempt Rate, measured in sessions-per-second (sps), which the DUT can process with zero errors over a defined, extended period. This period is defined as the amount of time needed to attempt N SIP sessions, where N is a parameter of test, at the attempt rate, R. An iterative process is used to find this rate. The algorithm corresponding to this process converges to R.

If the DUT vendor provides a value for R, the tester can use this value. In cases where the DUT vendor does not provide a value for R, or where the tester wants to establish the R of a system using local media characteristics, the algorithm should be run by setting "r", the session attempt rate, equal to a value of the tester's choice. For example the tester may initialize "r = 100" to start the algorithm and observe the value at convergence. The algorithm dynamically increases and decreases "r" as it converges to the a maximum sps value for R. The dynamic increase and decrease rate is controlled by the weights "w" and "d", respectively.

The pseudo-code corresponding to the description above follows, and a simulation written in the R statistical language is provided in Appendix A.

```
; ---- Parameters of test, adjust as needed
N  := 50000  ; Global maximum; once largest session rate has
             ; been established, send this many requests before
             ; calling the test a success
m  := {...}  ; Other attributes that affect testing, such
```

```
                 ; as media streams, etc.
r  := 100        ; Initial session attempt rate (in sessions/sec).
                 ; Adjust as needed (for example, if DUT can handle
                 ; thousands of calls in steady state, set to
                 ; appropriate value in the thousands).
w  := 0.10       ; Traffic increase weight (0 < w <= 1.0)
d  := max(0.10, w / 2)    ; Traffic decrease weight

; ---- End of parameters of test

proc find_R

   R = max_sps(r, m, N)  ; Setup r sps, each with m media
   ; characteristics until N sessions have been attempted.
   ; Note that if a DUT vendor provides this number, the tester
   ; can use the number as a Session Attempt Rate, R, instead
   ; of invoking max_sps()

end proc

; Iterative process to figure out the largest number of
; sps that we can achieve in order to setup n sessions.
; This function converges to R, the Session Attempt Rate.
proc max_sps(r, m, n)
   s     := 0    ; session setup rate
   old_r := 0    ; old session setup rate
   h     := 0    ; Return value, R
   count := 0

   ; Note that if w is small (say, 0.10) and r is small
   ; (say, <= 9), the algorithm will not converge since it
   ; uses floor() to increment r dynamically.  It is best
   ; off to start with the defaults (w = 0.10 and
   ; r >= 100)

   while (TRUE) {
      s := send_traffic(r, m, n) ; Send r sps, with m media
      ; characteristics until n sessions have been attempted.
      if (s == n)  {
          if (r > old_r)  {
              old_r = r
          }
          else  {
              count = count + 1
              if (count >= 10)  {
                  # We’ve converged.
                  h := max(r, old_r)
                  break
```

```
                       }
                   }

                   r  := floor(r + (w * r))
               }
               else  {
                   r := floor(r - (d * r))
                   d := max(0.10, d / 2)
                   w := max(0.10, w / 2)
               }

            }
            return h
         end proc
```

5. Reporting Format

5.1. Test Setup Report

```
SIP Transport Protocol = ___________________________
(valid values: TCP|UDP|TLS|SCTP|websockets|specify-other)
(specify if same transport used for connections to the DUT
and connections from the DUT.  If different transports
used on each connection, enumerate the transports used)

Connection management strategy for connection oriented
transports
   DUT receives requests on one connection = _______
   (Yes or no.  If no, DUT accepts a new connection for
   every incoming request, sends a response on that
   connection and closes the connection)
   DUT sends requests on one connection = __________
   (yes or no.  If no, DUT initiates a new connection to
   send out each request, gets a response on that
   connection and closes the connection)

Session Attempt Rate  _______________________________
(Session attempts/sec)
(The initial value for "r" in Benchmarking Algorithm of
Section 4.10)

Session Duration = _________________________________
(In seconds)
```

```
Total Sessions Attempted = _________________________
(Total sessions to be created over duration of test)

Media Streams Per Session =  _______________________
(number of streams per session)

Associated Media Protocol =  _______________________
(RTP|SRTP|specify-other)

Codec = ____________________________________________
(Codec type as identified by the organization that
specifies the codec)

Media Packet Size (audio only) =  __________________
(Number of bytes in an audio packet)

Establishment Threshold time =  ____________________
(Seconds)

TLS ciphersuite used
(for tests involving TLS) = ________________________
(E.g., TLS_RSA_WITH_AES_128_CBC_SHA)

IPSec profile used
(For tests involving IPSEC) = _____________________
```

5.2. Device Benchmarks for session setup

```
Session Establishment Rate, "R" = __________________
(sessions per second)
Is DUT acting as a media relay (yes/no) = _________
```

5.3. Device Benchmarks for registrations

```
Registration Rate =  ____________________________
(registrations per second)

Re-registration Rate =  ____________________________
(registrations per second)

Notes = ____________________________________________
(List any specific backend processing required or
other parameters that may impact the rate)
```

6. Test Cases

6.1. Baseline Session Establishment Rate of the test bed

Objective:
To benchmark the Session Establishment Rate of the Emulated Agent (EA) with zero failures.

Procedure:
1. Configure the DUT in the test topology shown in Figure 1.
2. Set media streams per session to 0.
3. Execute benchmarking algorithm as defined in Section 4.10 to get the baseline session establishment rate. This rate MUST be recorded using any pertinent parameters as shown in the reporting format of Section 5.1.

Expected Results: This is the scenario to obtain the maximum Session Establishment Rate of the EA and the test bed when no DUT is present. The results of this test might be used to normalize test results performed on different test beds or simply to better understand the impact of the DUT on the test bed in question.

6.2. Session Establishment Rate without media

Objective:
To benchmark the Session Establishment Rate of the DUT with no associated media and zero failures.

Procedure:
1. Configure a DUT according to the test topology shown in Figure 1 or Figure 2.
2. Set media streams per session to 0.
3. Execute benchmarking algorithm as defined in Section 4.10 to get the session establishment rate. This rate MUST be recorded using any pertinent parameters as shown in the reporting format of Section 5.1.

Expected Results: Find the Session Establishment Rate of the DUT when the EA is not sending media streams.

6.3. Session Establishment Rate with Media not on DUT

Objective:
To benchmark the Session Establishment Rate of the DUT with zero failures when Associated Media is included in the benchmark test but the media is not running through the DUT.

Procedure:

1. Configure a DUT according to the test topology shown in Figure 1.
2. Set media streams per session to 1.
3. Execute benchmarking algorithm as defined in Section 4.10 to get the session establishment rate with media. This rate MUST be recorded using any pertinent parameters as shown in the reporting format of Section 5.1.

Expected Results: Session Establishment Rate results obtained with Associated Media with any number of media streams per SIP session are expected to be identical to the Session Establishment Rate results obtained without media in the case where the DUT is running on a platform separate from the Media Relay.

6.4. Session Establishment Rate with Media on DUT

Objective:
To benchmark the Session Establishment Rate of the DUT with zero failures when Associated Media is included in the benchmark test and the media is running through the DUT.

Procedure:

1. Configure a DUT according to the test topology shown in Figure 2.
2. Set media streams per session to 1.
3. Execute benchmarking algorithm as defined in Section 4.10 to get the session establishment rate with media. This rate MUST be recorded using any pertinent parameters as shown in the reporting format of Section 5.1.

Expected Results: Session Establishment Rate results obtained with Associated Media may be lower than those obtained without media in the case where the DUT and the Media Relay are running on the same platform. It may be helpful for the tester to be aware of the reasons for this degradation, although these reasons are not parameters of the test. For example, the degree of performance degradation may be due to what the DUT does with the media (e.g., relaying vs. transcoding), the type of media (audio vs. video vs. data), and the codec used for the media. There may also be cases where there is no performance impact, if the DUT has dedicated media-path hardware.

6.5. Session Establishment Rate with TLS Encrypted SIP

Objective:
To benchmark the Session Establishment Rate of the DUT with zero failures when using TLS encrypted SIP signaling.

Procedure:

1. If the DUT is being benchmarked as a proxy or B2BUA, then configure the DUT in the test topology shown in Figure 1 or Figure 2.
2. Configure the tester to enable TLS over the transport being used during benchmarking. Note the ciphersuite being used for TLS and record it in Section 5.1.
3. Set media streams per session to 0 (media is not used in this test).
4. Execute benchmarking algorithm as defined in Section 4.10 to get the session establishment rate with TLS encryption.

Expected Results: Session Establishment Rate results obtained with TLS Encrypted SIP may be lower than those obtained with plaintext SIP.

6.6. Session Establishment Rate with IPsec Encrypted SIP

Objective:
To benchmark the Session Establishment Rate of the DUT with zero failures when using IPsec Encrypted SIP signaling.

Procedure:

1. Configure a DUT according to the test topology shown in Figure 1 or Figure 2.
2. Set media streams per session to 0 (media is not used in this test).
3. Configure tester for IPSec. Note the IPSec profile being used for and record it in Section 5.1.
4. Execute benchmarking algorithm as defined in Section 4.10 to get the session establishment rate with encryption.

Expected Results: Session Establishment Rate results obtained with IPSec Encrypted SIP may be lower than those obtained with plaintext SIP.

6.7. Registration Rate

Objective:
To benchmark the maximum registration rate the DUT can handle over an extended time period with zero failures.

Procedure:

1. Configure a DUT according to the test topology shown in Figure 3.
2. Set the registration timeout value to at least 3600 seconds.
3. Each register request MUST be made to a distinct address of record (AoR). Execute benchmarking algorithm as defined in Section 4.10 to get the maximum registration rate. This rate MUST be recorded using any pertinent parameters as shown in the reporting format of Section 5.1. For example, the use of TLS or IPSec during registration must be noted in the reporting format. In the same vein, any specific backend processing (use of databases, authentication servers, etc.) SHOULD be recorded as well.

Expected Results: Provides a maximum registration rate.

6.8. Re-Registration Rate

Objective:
To benchmark the re-registration rate of the DUT with zero failures using the same backend processing and parameters used during Section 6.7.

Procedure:

1. Configure a DUT according to the test topology shown in Figure 3.
2. First, execute test detailed in Section 6.7 to register the endpoints with the registrar and obtain the registration rate.
3. After at least 5 minutes of Step 2, but no more than 10 minutes after Step 2 has been performed, re-register the same AoRs used in Step 3 of Section 6.7. This will count as a re-registration because the SIP AoRs have not yet expired.

Expected Results: Note the rate obtained through this test for comparison with the rate obtained in Section 6.7.

7. IANA Considerations

This document does not requires any IANA considerations.

8. Security Considerations

Documents of this type do not directly affect the security of Internet or corporate networks as long as benchmarking is not performed on devices or systems connected to production networks.

Security threats and how to counter these in SIP and the media layer is discussed in RFC3261, RFC3550, and RFC3711 and various other drafts. This document attempts to formalize a set of common methodology for benchmarking performance of SIP devices in a lab environment.

9. Acknowledgments

The authors would like to thank Keith Drage and Daryl Malas for their contributions to this document. Dale Worley provided an extensive review that lead to improvements in the documents. We are grateful to Barry Constantine, William Cerveny and Robert Sparks for providing valuable comments during the document's last calls and expert reviews. Al Morton and Sarah Banks have been exemplary working group chairs, we thank them for tracking this work to completion. Tom Taylor provided an in-depth review and subsequent comments on the benchmarking convergence algorithm in Section 4.10.

10. References

10.1. Normative References

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

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

[I-D.sip-bench-term] Davids, C., Gurbani, V., and S. Poretsky, "SIP Performance Benchmarking Terminology", draft-ietf-bmwg-sip-bench-term-12 (work in progress), November 2014.

10.2. Informative References

[RFC3261] Rosenberg, J., Schulzrinne, H., Camarillo, G., Johnston, A., Peterson, J., Sparks, R., Handley, M., and E. Schooler, "SIP: Session Initiation Protocol", RFC 3261, June 2002.

[Rtool] R Development Core Team, "R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/", , 2011.

Appendix A. R Code Component to simulate benchmarking algorithm

```
# Copyright (c) 2014 IETF Trust and Vijay K. Gurbani.  All
# rights reserved.
#
# Redistribution and use in source and binary forms, with
# or without modification, are permitted provided that the
# following conditions are met:
#
# * Redistributions of source code must retain the above
#   copyright notice, this list of conditions and the following
#   disclaimer.
# * Redistributions in binary form must reproduce the above
#   copyright notice, this list of conditions and the following
#   disclaimer in the documentation and/or other materials
#   provided with the distribution.
# * Neither the name of Internet Society, IETF or IETF Trust,
#   nor the names of specific contributors, may be used
#   to endorse or promote products derived from this software
#   without specific prior written permission.
#
# THIS SOFTWARE IS PROVIDED BY THE COPYRIGHT HOLDERS AND
# CONTRIBUTORS "AS IS" AND ANY EXPRESS OR IMPLIED WARRANTIES,
# INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF
# MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE ARE
# DISCLAIMED. IN NO EVENT SHALL THE COPYRIGHT OWNER OR
# CONTRIBUTORS BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL,
# SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING,
# BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR
# SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS
# INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY,
# WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING
# NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE
# USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY
# OF SUCH DAMAGE.

w = 0.10
d = max(0.10, w / 2)
DUT_max_sps = 460     # Change as needed to set the max sps value
                      # for a DUT

# Returns R, given r (initial session attempt rate).
# E.g., assume that a DUT handles 460 sps in steady state
# and you have saved this code in a file simulate.r.  Then,
# start an R session and do the following:
#
# > source("simulate.r")
```

```
      # > find_R(100)
      # ... debug output omitted ...
      # [1] 458
      #
      # Thus, the max sps that the DUT can handle is 458 sps, which is
      # close to the absolute maximum of 460 sps the DUT is specified to
      # do.
      find_R <- function(r)  {
         s     = 0
         old_r = 0
         h     = 0
         count = 0

         # Note that if w is small (say, 0.10) and r is small
         # (say, <= 9), the algorithm will not converge since it
         # uses floor() to increment r dynamically.  It is best
         # off to start with the defaults (w = 0.10 and
         # r >= 100)

         cat("r   old_r    w     d \n")
         while (TRUE)  {
            cat(r, ' ', old_r, ' ', w, ' ', d, '\n')
            s = send_traffic(r)
            if (s == TRUE)  {     # All sessions succeeded

                if (r > old_r)  {
                    old_r = r
                }
                else  {
                    count = count + 1

                    if (count >= 10)  {
                        # We've converged.
                        h = max(r, old_r)
                        break
                    }
                }

                r  = floor(r + (w * r))
            }
            else  {
                r = floor(r - (d * r))
                d = max(0.10, d / 2)
                w = max(0.10, w / 2)
            }
         }

         h
```

```
      }

      send_traffic <- function(r)  {
         n = TRUE

         if (r > DUT_max_sps)  {
             n = FALSE
         }

         n
      }
```

Authors' Addresses

Carol Davids
Illinois Institute of Technology
201 East Loop Road
Wheaton, IL 60187
USA

Phone: +1 630 682 6024
Email: davids@iit.edu

Vijay K. Gurbani
Bell Laboratories, Alcatel-Lucent
1960 Lucent Lane
Rm 9C-533
Naperville, IL 60566
USA

Phone: +1 630 224 0216
Email: vkg@bell-labs.com

Scott Poretsky
Allot Communications
300 TradeCenter, Suite 4680
Woburn, MA 08101
USA

Phone: +1 508 309 2179
Email: sporetsky@allot.com

Benchmarking Methodology Working Group C. Davids
Internet-Draft Illinois Institute of Technology
Intended status: Informational V. Gurbani
Expires: May 16, 2015 Bell Laboratories,
Alcatel-Lucent
S. Poretsky
Allot Communications
November 12, 2014

Terminology for Benchmarking Session Initiation Protocol (SIP) Devices: Basic session setup and registration
draft-ietf-bmwg-sip-bench-term-12

Abstract

This document provides a terminology for benchmarking the Session Initiation Protocol (SIP) performance of devices. Methodology related to benchmarking SIP devices is described in the companion methodology document. Using these two documents, benchmarks can be obtained and compared for different types of devices such as SIP Proxy Servers, Registrars and Session Border Controllers. The term "performance" in this context means the capacity of the device-under-test (DUT) to process SIP messages. Media streams are used only to study how they impact the signaling behavior. The intent of the two documents is to provide a normalized set of tests that will enable an objective comparison of the capacity of SIP devices. Test setup parameters and a methodology is necessary because SIP allows a wide range of configuration and operational conditions that can influence performance benchmark measurements. A standard terminology and methodology will ensure that benchmarks have consistent definition and were obtained following the same procedures.

Status of this Memo

This Internet-Draft will expire on May 16, 2015.

Copyright Notice

Table of Contents

1. Terminology

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 BCP 14, RFC2119 [RFC2119]. RFC 2119 defines the use of these key words to help make the intent of standards track documents as clear as possible. While this document uses these keywords, this document is not a standards track document. The term Throughput is defined in RFC2544 [RFC2544].

For the sake of clarity and continuity, this document adopts the template for definitions set out in Section 2 of RFC 1242 [RFC1242].

The term Device Under Test (DUT) is defined in the following BMWG documents:

Device Under Test (DUT) (c.f., Section 3.1.1 RFC 2285 [RFC2285]).

Many commonly used SIP terms in this document are defined in RFC 3261 [RFC3261]. For convenience the most important of these are reproduced below. Use of these terms in this document is consistent with their corresponding definition in the base SIP specification [RFC3261] as amended by [RFC4320], [RFC5393] and [RFC6026].

- Call Stateful: A proxy is call stateful if it retains state for a dialog from the initiating INVITE to the terminating BYE request. A call stateful proxy is always transaction stateful, but the converse is not necessarily true.
- Stateful Proxy: A logical entity, as defined by [RFC3261], that maintains the client and server transaction state machines during the processing of a request. (Also known as a transaction stateful proxy.) The behavior of a stateful proxy is further defined in Section 16 of RFC 3261 [RFC3261] . A transaction stateful proxy is not the same as a call stateful proxy.
- Back-to-back User Agent: A back-to-back user agent (B2BUA) is a logical entity that receives a request and processes it as a user agent server (UAS). In order to determine how the request should be answered, it acts as a user agent client (UAC) and generates requests. Unlike a proxy server, it maintains dialog state and must participate in all requests sent on the dialogues it has established. Since it is a concatenation of a UAC and a UAS, no explicit definitions are needed for its behavior.

2. Introduction

Service Providers and IT Organizations deliver Voice Over IP (VoIP) and Multimedia network services based on the IETF Session Initiation

Protocol (SIP) [RFC3261]. SIP is a signaling protocol originally intended to be used to dynamically establish, disconnect and modify streams of media between end users. As it has evolved it has been adopted for use in a growing number of services and applications. Many of these result in the creation of a media session, but some do not. Examples of this latter group include text messaging and subscription services. The set of benchmarking terms provided in this document is intended for use with any SIP-enabled device performing SIP functions in the interior of the network, whether or not these result in the creation of media sessions. The performance of end-user devices is outside the scope of this document.

A number of networking devices have been developed to support SIP-based VoIP services. These include SIP Servers, Session Border Controllers (SBC) and Back-to-back User Agents (B2BUA). These devices contain a mix of voice and IP functions whose performance may be reported using metrics defined by the equipment manufacturer or vendor. The Service Provider or IT Organization seeking to compare the performance of such devices will not be able to do so using these vendor-specific metrics, whose conditions of test and algorithms for collection are often unspecified.

SIP functional elements and the devices that include them can be configured many different ways and can be organized into various topologies. These configuration and topological choices impact the value of any chosen signaling benchmark. Unless these conditions-of-test are defined, a true comparison of performance metrics across multiple vendor implementations will not be possible.

Some SIP-enabled devices terminate or relay media as well as signaling. The processing of media by the device impacts the signaling performance. As a result, the conditions-of-test must include information as to whether or not the device under test processes media. If the device processes media during the test, a description of the media must be provided. This document and its companion methodology document [I-D.ietf-bmwg-sip-bench-meth] provide a set of black-box benchmarks for describing and comparing the performance of devices that incorporate the SIP User Agent Client and Server functions and that operate in the network's core.

The definition of SIP performance benchmarks necessarily includes definitions of Test Setup Parameters and a test methodology. These enable the Tester to perform benchmarking tests on different devices and to achieve comparable results. This document provides a common set of definitions for Test Components, Test Setup Parameters, and Benchmarks. All the benchmarks defined are black-box measurements of the SIP signaling plane. The Test Setup Parameters and Benchmarks defined in this document are intended for use with the companion

Methodology document.

2.1. Scope

The scope of this document is summarized as follows:

- o This terminology document describes SIP signaling performance benchmarks for black-box measurements of SIP networking devices. Stress and debug scenarios are not addressed in this document.
- o The DUT must be RFC 3261 capable network equipment. This may be a Registrar, Redirect Server, or Stateful Proxy. This document does not require the intermediary to assume the role of a stateless proxy. A DUT may also include a B2BUA, SBC functionality.
- o The Tester acts as multiple "Emulated Agents" (EA) that initiate (or respond to) SIP messages as session endpoints and source (or receive) associated media for established connections.
- o SIP Signaling in presence of media
    - * The media performance is not benchmarked.
    - * Some tests require media, but the use of media is limited to observing the performance of SIP signaling. Tests that require media will annotate the media characteristics as a condition of test.
    - * The type of DUT dictates whether the associated media streams traverse the DUT. Both scenarios are within the scope of this document.
    - * SIP is frequently used to create media streams; the signaling plane and media plane are treated as orthogonal to each other in this document. While many devices support the creation of media streams, benchmarks that measure the performance of these streams are outside the scope of this document and its companion methodology document [I-D.ietf-bmwg-sip-bench-meth]. Tests may be performed with or without the creation of media streams. The presence or absence of media streams MUST be noted as a condition of the test as the performance of SIP devices may vary accordingly. Even if the media is used during benchmarking, only the SIP performance will be benchmarked, not the media performance or quality.
- o Both INVITE and non-INVITE scenarios (registrations) are addressed in this document. However, benchmarking SIP presence or subscribe-notify extensions is not a part of this document.
- o Different transport -- such as UDP, TCP, SCTP, or TLS -- may be used. The specific transport mechanism MUST be noted as a condition of the test as the performance of SIP devices may vary accordingly.
- o REGISTER and INVITE requests may be challenged or remain unchallenged for authentication purpose. Whether or not the REGISTER and INVITE requests are challenged is a condition of test which will be recorded along with other such parameters which may impact the SIP performance of the device or system under test.

- o Re-INVITE requests are not considered in scope of this document since the benchmarks for INVITEs are based on the dialog created by the INVITE and not on the transactions that take place within that dialog.
- o Only session establishment is considered for the performance benchmarks. Session disconnect is not considered in the scope of this document. This is because our goal is to determine the maximum capacity of the device or system under test, that is the number of simultaneous SIP sessions that the device or system can support. It is true that there are BYE requests being created during the test process. These transactions do contribute to the load on the device or system under test and thus are accounted for in the metric we derive. We do not seek a separate metric for the number of BYE transactions a device or system can support.
- o IMS-specific scenarios are not considered, but test cases can be applied with 3GPP-specific SIP signaling and the P-CSCF as a DUT.
- o The benchmarks described in this document are intended for a laboratory environment and are not intended to be used on a production network. Some of the benchmarks send enough traffic that a denial of service attack is possible if used in production networks.

3. Term Definitions

3.1. Protocol Components

3.1.1. Session

Definition:
The combination of signaling and media messages and associated processing that enable a single SIP-based audio or video call, or SIP registration.

Discussion:
The term "session" commonly implies a media session. In this document the term is extended to cover the signaling and any media specified and invoked by the corresponding signaling.

Measurement Units:
N/A.

Issues:
None.

See Also:
Media Plane
Signaling Plane
Associated Media

3.1.2. Signaling Plane

Definition:
The plane in which SIP messages [RFC3261] are exchanged between SIP Agents [RFC3261].

Discussion:
SIP messages are used to establish sessions in several ways: directly between two User Agents [RFC3261], through a Proxy Server [RFC3261], or through a series of Proxy Servers. The Session Description Protocol (SDP) is included in the Signaling Plane.

Measurement Units:
N/A.

Issues:
None.

See Also:
Media Plane
EAs

3.1.3. Media Plane

Definition:
The data plane in which one or more media streams and their associated media control protocols (e.g., RTCP [RFC3550]) are exchanged between User Agents after a media connection has been created by the exchange of signaling messages in the Signaling Plane.

Discussion:
Media may also be known as the "bearer channel". The Media Plane MUST include the media control protocol, if one is used, and the media stream(s). Examples of media are audio and video. The media streams are described in the SDP of the Signaling Plane.

Measurement Units:

N/A.

Issues:
None.

See Also:
Signaling Plane

3.1.4. Associated Media

Definition:
Media that corresponds to an 'm' line in the SDP payload of the Signaling Plane.

Discussion:
The format of the media is determined by the SDP attributes for the corresponding 'm' line.

Measurement Units:
N/A.

Issues:
None.

3.1.5. Overload

Definition:
Overload is defined as the state where a SIP server does not have sufficient resources to process all incoming SIP messages [RFC6357].

Discussion:
The distinction between an overload condition and other failure scenarios is outside the scope of black box testing and of this document. Under overload conditions, all or a percentage of Session Attempts will fail due to lack of resources. In black box testing the cause of the failure is not explored. The fact that a failure occurred for whatever reason, will trigger the tester to reduce the offered load, as described in the companion methodology document, [I-D.ietf-bmwg-sip-bench-meth]. SIP server resources may include CPU processing capacity, network bandwidth, input/output queues, or disk resources. Any combination of resources may be fully utilized when a SIP server (the DUT) is in the overload condition. For proxy-only (or intermediary) devices, it is expected that the proxy will be driven into overload based on the delivery rate of signaling requests.

Measurement Units:
N/A.

3.1.6. Session Attempt

Definition:
A SIP INVITE or REGISTER request sent by the EA that has not received a final response.

Discussion:
The attempted session may be either an invitation to an audio/video communication or a registration attempt. When counting the number of session attempts we include all requests that are rejected for lack of authentication information. The EA needs to record the total number of session attempts including those attempts that are routinely rejected by a proxy that requires the UA to authenticate itself. The EA is provisioned to deliver a specific number of session attempts per second. But the EA must also count the actual number of session attempts per given time interval.

Measurement Units:
N/A.

Issues:
None.

See Also:
Session
Session Attempt Rate

3.1.7. Established Session

Definition:
A SIP session for which the EA acting as the UE/UA has received a 200 OK message.

Discussion:
An Established Session may be either an invitation to an audio/video communication or a registration attempt. Early dialogues for INVITE requests are out of scope for this work.

Measurement Units:

N/A.

Issues:
None.

See Also:
None.

3.1.8. Session Attempt Failure

Definition:
A session attempt that does not result in an Established Session.

Discussion:
The session attempt failure may be indicated by the following observations at the EA:

1. Receipt of a SIP 3xx-, 4xx-, 5xx-, or 6xx-class response to a Session Attempt.
2. The lack of any received SIP response to a Session Attempt within the Establishment Threshold Time (c.f. Section 3.3.2).

Measurement Units:
N/A.

Issues:
None.

See Also:
Session Attempt

3.2. Test Components

3.2.1. Emulated Agent

Definition:
A device in the test topology that initiates/responds to SIP messages as one or more session endpoints and, wherever applicable, sources/receives Associated Media for Established Sessions.

Discussion:
The EA functions in the Signaling and Media Planes. The Tester may act as multiple EAs.

Measurement Units:
N/A

Issues:
None.

See Also:
Media Plane
Signaling Plane
Established Session
Associated Media

3.2.2. Signaling Server

Definition:
Device in the test topology that facilitates the creation of sessions between EAs. This device is the DUT.

Discussion:
The DUT is a RFC3261-capable network intermediary such as a Registrar, Redirect Server, Stateful Proxy, B2BUA or SBC.

Measurement Units:
NA

Issues:
None.

See Also:
Signaling Plane

3.2.3. SIP Transport Protocol

Definition:
The protocol used for transport of the Signaling Plane messages.

Discussion:
Performance benchmarks may vary for the same SIP networking device depending upon whether TCP, UDP, TLS, SCTP, websockets [RFC7118] or any future transport layer protocol is used. For this reason it is necessary to measure the SIP Performance Benchmarks using these various transport protocols. Performance Benchmarks MUST report the SIP Transport Protocol used to obtain the benchmark results.

Measurement Units:
While these are not units of measure, they are attributes that are one of many factors that will contribute to the value of the measurements to be taken. TCP, UDP, SCTP, TLS over TCP, TLS over UDP, TLS over SCTP, and websockets are among the possible values to be recorded as part of the test.

Issues:
None.

See Also:
None.

3.3. Test Setup Parameters

3.3.1. Session Attempt Rate

Definition:
Configuration of the EA for the number of sessions per second (sps) that the EA attempts to establish using the services of the DUT.

Discussion:
The Session Attempt Rate is the number of sessions per second that the EA sends toward the DUT. Some of the sessions attempted may not result in a session being established.

Measurement Units:
Session attempts per second

Issues:
None.

See Also:
Session
Session Attempt

3.3.2. Establishment Threshold Time

Definition:
Configuration of the EA that represents the amount of time that an EA client will wait for a response from an EA server before declaring a Session Attempt Failure.

Discussion:
This time duration is test dependent.

It is RECOMMENDED that the Establishment Threshold Time value be set to Timer B or Timer F as specified in RFC 3261, Table 4 [RFC3261].

Measurement Units:
Seconds

Issues:
None.

See Also:
None.

3.3.3. Session Duration

Definition:
Configuration of the EA that represents the amount of time that the SIP dialog is intended to exist between the two EAs associated with the test.

Discussion:
The time at which the BYE is sent will control the Session Duration.

Measurement Units:
seconds

Issues:
None.

See Also:
None.

3.3.4. Media Packet Size

Definition:
Configuration on the EA for a fixed number of frames or samples to be sent in each RTP packet of the media stream when the test involves Associated Media.

Discussion:
This document describes a method to measure SIP performance. If the DUT is processing media as well as SIP messages the media processing will potentially slow down the SIP processing and lower the SIP performance metric. The tests with associated media are designed for audio codecs and the assumption was made that larger media packets would require more processor time. This document does not define parameters applicable to video codecs.

For a single benchmark test, media sessions use a defined number of samples or frames per RTP packet. If two SBCs, for example, used the same codec but one puts more frames into the RTP packet, this might cause variation in the performance benchmark results.

Measurement Units:
An integer number of frames or samples, depending on whether hybrid- or sample-based codec are used, respectively.

Issues:
None.

See Also:
None.

3.3.5. Codec Type

Definition:
The name of the codec used to generate the media session.

Discussion
For a single benchmark test, all sessions use the same size packet for media streams. The size of packets can cause a variation in the performance benchmark measurements.

Measurement Units:
This is a textual name (alphanumeric) assigned to uniquely identify the codec.

Issues:
None.
See Also:
None.

3.4. Benchmarks

3.4.1. Session Establishment Rate

Definition:
The maximum value of the Session Attempt Rate that the DUT can handle for an extended, pre-defined, period with zero failures.

Discussion:
This benchmark is obtained with zero failure. The session attempt rate provisioned on the EA is raised and lowered as described in the algorithm in the accompanying methodology document [I-D.ietf-bmwg-sip-bench-meth], until a traffic load over the period of time necessary to attempt N sessions completes without failure, where N is a parameter specified in the algorithm and recorded in the Test Setup Report.

Measurement Units:
sessions per second (sps)

Issues:
None.

See Also:
Invite-Initiated Sessions
Non-Invite-Initiated Sessions
Session Attempt Rate

3.4.2. Registration Rate

Definition:
The maximum value of the Registration Attempt Rate that the DUT can handle for an extended, pre-defined, period with zero failures.

Discussion:
This benchmark is obtained with zero failures. The registration rate provisioned on the Emulated Agent is raised and lowered as described in the algorithm in the companion methodology draft [I-D.ietf-bmwg-sip-bench-meth], until a traffic load consisting of registration attempts at the given attempt rate over the period of time necessary to attempt N registrations completes without failure, where N is a parameter specified in the algorithm and recorded in the Test Setup Report.
This benchmark is described separately from the Session Establishment Rate (Section 3.4.1), although it could be considered a special case of that benchmark, since a REGISTER request is a request for a Non-Invite-Initiated session. It is defined separately because it is a very important benchmark for most SIP installations. An example demonstrating its use is an

avalanche restart, where hundreds of thousands of end points register simultaneously following a power outage. In such a case, an authoritative measurement of the capacity of the device to register endpoints is useful to the network designer. Additionally, in certain controlled networks, there appears to be a difference between the registration rate of new endpoints and the registering rate of existing endpoints (register refreshes). This benchmark can capture these differences as well.

Measurement Units:
registrations per second (rps)

Issues:
None.

See Also:
None.

3.4.3. Registration Attempt Rate

Definition:
Configuration of the EA for the number of registrations per second that the EA attempts to send to the DUT.

Discussion:
The Registration Attempt Rate is the number of registration requests per second that the EA sends toward the DUT.

Measurement Units:
Registrations per second (rps)

Issues:
None.

See Also: Non-Invite-Initiated Session

4. IANA Considerations

This document requires no IANA considerations.

5. Security Considerations

Documents of this type do not directly affect the security of Internet or corporate networks as long as benchmarking is not performed on devices or systems connected to production networks. Security threats and how to counter these in SIP and the media layer is discussed in RFC3261 [RFC3261], RFC 3550 [RFC3550] and RFC3711 [RFC3711]. This document attempts to formalize a set of common terminology for benchmarking SIP networks. Packets with unintended and/or unauthorized DSCP or IP precedence values may present security issues. Determining the security consequences of such packets is out of scope for this document.

6. Acknowledgments

The authors would like to thank Keith Drage, Cullen Jennings, Daryl Malas, Al Morton, and Henning Schulzrinne for invaluable contributions to this document. Dale Worley provided an extensive review that lead to improvements in the documents. We are grateful to Barry Constantine, William Cerveny and Robert Sparks for providing valuable comments during the document's last calls and expert reviews. Al Morton and Sarah Banks have been exemplary working group chairs, we thank them for tracking this work to completion.

7. References

7.1. Normative References

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

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

[RFC3261] Rosenberg, J., Schulzrinne, H., Camarillo, G., Johnston, A., Peterson, J., Sparks, R., Handley, M., and E. Schooler, "SIP: Session Initiation Protocol", RFC 3261, June 2002.

[RFC5393] Sparks, R., Lawrence, S., Hawrylyshen, A., and B. Campen, "Addressing an Amplification Vulnerability in Session Initiation Protocol (SIP) Forking Proxies", RFC 5393, December 2008.

[RFC4320] Sparks, R., "Actions Addressing Identified Issues with the Session Initiation Protocol's (SIP) Non-INVITE Transaction", RFC 4320, January 2006.

[RFC6026] Sparks, R. and T. Zourzouvillys, "Correct Transaction Handling for 2xx Responses to Session Initiation Protocol (SIP) INVITE Requests", RFC 6026, September 2010.

[I-D.ietf-bmwg-sip-bench-meth]
Davids, C., Gurbani, V., and S. Poretsky, "SIP Performance Benchmarking Methodology", draft-ietf-bmwg-sip-bench-meth-10 (work in progress), May 2014.

7.2. Informational References

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

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

[RFC3550] Schulzrinne, H., Casner, S., Frederick, R., and V. Jacobson, "RTP: A Transport Protocol for Real-Time Applications", STD 64, RFC 3550, July 2003.

[RFC3711] Baugher, M., McGrew, D., Naslund, M., Carrara, E., and K. Norrman, "The Secure Real-time Transport Protocol (SRTP)", RFC 3711, March 2004.

[RFC6357] Hilt, V., Noel, E., Shen, C., and A. Abdelal, "Design Considerations for Session Initiation Protocol (SIP) Overload Control", RFC 6357, August 2011.

[RFC7118] Baz Castillo, I., Millan Villegas, J., and V. Pascual, "The Websocket Protocol as a Transport for the Session Initiation Protocol (SIP)", RFC 7118, January 2014.

Authors' Addresses

Carol Davids
Illinois Institute of Technology
201 East Loop Road
Wheaton, IL 60187
USA

Phone: +1 630 682 6024
Email: davids@iit.edu

Vijay K. Gurbani
Bell Laboratories, Alcatel-Lucent
1960 Lucent Lane
Rm 9C-533
Naperville, IL 60566
USA

Phone: +1 630 224 0216
Email: vkg@bell-labs.com

Scott Poretsky
Allot Communications
300 TradeCenter, Suite 4680
Woburn, MA 08101
USA

Phone: +1 508 309 2179
Email: sporetsky@allot.com

Network Working Group V. Manral, Ed.
Internet-Draft IPInfusion Inc.
Intended status: Informational May 23, 2010
Expires: November 24, 2010

Benchmarking Power usage of networking devices
draft-manral-bmwg-power-usage-01

Abstract

With the rapid growth of networks around the globe there is an ever increasing need to improve the energy efficiency of devices. Operators begining to seek more information of power consumption in the network, have no standard mechanism to measure, report and compare power usage of different networking equipment under different network configuration and conditions exist.

This document provides suggestions for measuring power usage of live networks under different traffic loads and various switch router configuration settings. It provides a suite which can be deployed on any networking device .

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].

Table of Contents

1. Introduction

Energy Efficiency is becoming increasing important in the operation of network infrastructure. Data traffic is exploding at an accelerated rate. Networks provide communication channels that facilitates components of the infrastructues to exchange critical information and are always on. On the other hand, a lot of devices run at very low average utlization rates. Various strategies are being defined to improve network utilization of these devices and thus improve power consumption.

The first step to obtain a network wide view is to start with an individual device view of the system and address different devices in the network on a per device basis. The easiest way to measure the power consumption of a device is to use a power meter. This can be used to measure power under a variety of conditions affecting power usage on a networking device.

Various techniques have been defined for energy management of networking devices. However, there is no common strategy to actually benchmark power utilization of networking devices like routers or switches. This document defines the mechanism to correctly characterize and benchmark the power consumption of various networking devices so as to be able to correctly measure and compare the power usage of various devices. This will enable intelligent decisions to optimize the power consumption for individual devices and the network as a whole. Benchmark are also required to compare effectiveness of various energy optimization techniques.

The Network Energy Consumption Rate (NECR) as well as Network Energy Proportionality Index (NEPI) is also defined here.

The procedures/ metrics defined in this document have been used to perform live measurement with a variety of networking equipment from three large well known vendors.

2. Challenges in defining benchmarks

Using the "Maximum Rated Power" and spec sheets of devices and adding the values for all devices are of little use because the measurement gives the maximum power that can consumed by the device, however that does not accurately reflect the power consumed by the device under a normal work load. Typical energy requirements of a networking device are dependent on device configuration and traffic.

The ratio of the actual power consumed by the device on an average, to its maximum rated power varies widely across different device families. Thus, relying merely on the maximum rated power can grossly overestimate the total energy consumed by networking equipment.

There are a wide variety of networking equipment and finding a general benchmark to work across a variety of devices, requires a lot of flexibility in benchmarking methodology. the workload and test conditions will also depend on the kind of device.

A network device consists of a lot of individual component, each of which consume power. For example, only considering the power consumption of the CPU/ data forwarding ASIC we may ignore the power consumption of the other components like external memory.

Power instrumentation of a device in a live network involves unplugging the device and plugging it into a power meter. This can inturn lead to traffic loss. Unfortunately, most current equipment is not equipped with internal instrumentation to report power usage of the device or its components. It is for this reason the power measurement is done on an individual device under different network conditions using a traffic generator.

The network devices can also dissipate significant heat. Past studies have shown dissipation rations of 2.5. Which means if the power in is 2.5 Watt, only 1 Watt is used for actual work, the rest is disspated as heat. This heating can lead to more power consumed by fan/ compressor for cooling the devices. Though this methodology does not measure the power consumed by external cooling infrastructure, it measures the power consumed internally. It also (optionally) measures the temperature change of the device which can be correlated to the amount of external power consumed to cool the device.

The amount of power used at startup can be more than the average power usage of the device. This is also measured as part of the test methodology.

3. Factors for power consumption

The metrics defined here will help operators get a more accurate idea of power consumed by network equipment and hence forecast their power budget. These will also help device vendors test and compare the new power efficiency enhancements on various devices.

3.1. Network Factors affecting power consumption

The first and the most important factor from the network perspective which can determine the power consumption is the traffic load. Benchmarks must be performed with different traffic loads in the network.

There are now various kinds of transcivers/ connectors on a network device. For the same bandwidth the power usage of a device depends on the kind of connector used. The connector/ interface type used needs to be specified in the benchmark.

The length of the cable used also defines the amount of power consumed by the system. Benchmarks should specify the cable length used. For example, a 5 meter cable can be used wherever possible.

3.2. Device Factors affecting power consumption

Base Chassis Power - typically, higher end network devices come with a chassis and card slots. Each slot may have a number of ports. For the lower end devices there are no removable card slots. In both these cases the base chassis power consists of processors, fans, memory, etc.

Number of line cards - In switches that support inserting linecards, there is a limit on the number of ports per linecard as well as the aggregate bandwidth that each linecard can accommodate. This mechanism allows network operators the flexibility to only plug in as many linecards as they need. For each benchmark the total number of line cards plugged into the system needs to be specified.

Number of active ports - This term refers to the total number of ports on the switch (across all the linecards) that are active (with cables plugged in). The remaining ports on the switch are explicitly disabled using the switchs command line interface. For each benchmark the number of active and passive ports must be specified.

Port settings - Setting this parameter limits the line rate forwarding capacity of individual ports. For each benchmark the port configuration and settings need to be specified.

Port Utilization - This term describes the actual throughput flowing through a port relative to its specified capacity. For each benchmark the port utilization of each port must be specified. The actual traffic can use the information defined in RFC 2544 [RFC2544].

TCAM - Network vendors typically implement packet classification in hardware. TCAMs are supported by most vendors as they have very fast look-up times. However, they are are notoriously power-hungry. The size of the TCAM in a switch is widely variable. The size of the TCAM needs to be reported in the benchmark document. The number of TCAM entries does not affect power consumption.

Firmware - Vendors periodically release upgraded versions of their switch/router firmware. Different versions of firmware may also impact the device power consumption. The firmware version needs to be reported in the benchmark document. Different firmware versions have resulted in different power usage.

3.3. Traffic Factors affecting power consumption

Packet Size - Different packet sizes typically do not effect power consumption.

Inter-Packet Delay - time between successive packets may affect power usage but we do not measure the effects in detail.

CPU traffic - Percentage of CPU traffic. For our benchmarks we can assume different values of CPU bound traffic. The different percentage of CPU bound traffic must be specified in the benchmark.

4. Network Energy Consumption Rate (NECR)

To optimize the run time energy usage for different devices, the additional energy consumption that will result as a factor of additional traffic needs to be known. The NECR defines the power usage increase in MilliWatts per Mbps of data at the physical layer.

The NECR will depend on the line card, the port and the other factors defined earlier.

For the effective use of the NECR the base power of the chassis, a line card and a port needs to be specified when there is no load. The measurements must take into consideration power optimization techniques when there is no traffic on any port of a line card.

5. Network Energy Proportionality Index (NEPI)

In the ideal case the power consumed by a device is proportional to its network load. The average difference between the ideal(I) and the measured (M) power consumption defines the EPI.

The ideal power is measured by assuming the power consumed by a device at 100% traffic load and using that to derive the ideal power usage for different traffic loads.

EPIx = (Mx - Ix)/ Mx * 100

EPI = EPI1 + EPI2 + ... EPIn / n

The EPI is independent of the actualy traffic load. It can thus be used to define the energy efficiency of a networking device. A value of 0 means the power usage is agnostic to traffic and a value of 100 means that the device has perfect energy proportionality.

6. Benchmark details

All power measurements are done in MilliWatts, except NECR which is done in MilliWatts/ Mbps.

7. IANA Considerations

This document makes no request of IANA.

Note to RFC Editor: this section may be removed on publication as an RFC.

8. Security Considerations

This document raises no new security issues.

9. Acknowledgements

This document derives a lot of its text and content from "A Power Benchmarking Framework for Network Devices" paper and the authors of that are duly acknowledged.

The author would like to thank Srini Seetharaman (srini.seetharaman@telekom.com) and Priya Mahadevan (priya.mahadevan@hp.com) for their support with the draft. The author would also like to thank Al Morton (AT&T) and Robert Peglar(XioTech) for his careful reading and suggestions on the draft.

10. References

10.1. Normative References

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

10.2. Informative References

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

Author's Address

Vishwas Manral (editor)
IPInfusion Inc.
1188 E. Arques Ave.
Sunnyvale, CA 94085
US

Phone: 408-400-1900
Fax:
Email: vishwas@ipinfusion.com
URI:

Network Working Group   A. Morton
Internet-Draft   AT&T Labs
Intended status: Informational   October 6, 2010
Expires: April 9, 2011

IMIX Genome: Specification of variable packet sizes for additional testing
draft-morton-bmwg-imix-genome-00

Abstract

Benchmarking Methodologies have always relied on test conditions with constant packet sizes, with the goal of understanding what network device capability has been tested. Constant packets sizes differ significantly from the conditions encountered in operational deployment, and so additional tests are sometimes conducted with a mixture of packet sizes, or "IMIX". The mixture of sizes a networking device will encounter is highly variable and depends on many factors. An IMIX suited for one networking device and deployment will not be appropriate for another. However, the mix of sizes may be known and the tester may be asked to augment the fixed size tests. To address this need, and the additional goal of repeatable test conditions, this draft proposes a way to specify the exact repeating sequence of packet sizes from the usual set of fixed sizes.

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].

Status of this Memo

This Internet-Draft will expire on April 9, 2011.

Table of Contents

1. Introduction

This memo defines a method to unambiguously specify the sequence of packet sizes used in a load test.

Benchmarking Methodologies [RFC2544] have always relied on test conditions with constant packet 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.

Constant packets sizes differ significantly from the conditions encountered in operational deployment, and so additional tests are sometimes conducted with a mixture of packet sizes. The set of sizes used is often called an Internet Mix, or "IMIX" [Spirent], [IXIA], [Agilent].

The mixture of sizes a networking device will encounter is highly variable and depends on many factors. An IMIX suited for one networking device and deployment will not be appropriate for another. However, the mix of sizes may be known and the tester may be asked to augment the fixed size tests.

To address this need, and the additional goal of repeatable test conditions, this draft proposes a way to specify the exact repeating sequence of packet sizes from the usual set of fixed sizes: the IMIX Genome.

1.1. First Draft

In this first draft, some section are very short or to-be-provided (TBP), and there are several questions identified for further discussion.

2. Scope and Goals

This memo defines a method to unambiguously specify the sequence of packet sizes that have been used in a load test, assuming that a relevant mix of sizes is known to the tester and the length of the repeating sequence is not very long (<30 packets).

The IMIX Genome will allow an exact sequence of packet sizes to be communicated as a single-line name, resolving the current ambiguity with results that simply refer to "IMIX".

While documentation of the exact sequence is ideal, the memo also covers the case where the sequence of sizes is very long or may be generated by a pseudo-random process.

It is a colossal non-goal to standardize one or more versions of the IMIX. This topic has been discussed on many occasions on the bmwg-list[IMIXonList]. The goal is to enable customization with minimal constraints while fostering repeatable testing once the fixed size testing is complete.

3. Specification of the IMIX Genome

The IMIX Genome is specified in the following format:

IMIX - 123456...x

where each number is replaced by the letter corresponding to the packet size of the packet at that position in the sequence. The following table gives the letter encoding for the [RFC2544] standard sizes (64, 128, 256, 512, 1024, 1280, and 1518 bytes).

Size, bytes	Genome Code Letter
64	a
128	b
256	c
512	d
1024	e
1280	f
1518	g
MTU ??	h

For example: a five packet sequence with sizes 64,64,64,1280,1518 would be designated:

IMIX - aaafg

While this approach allows some flexibility, there are also constraints.

- o Non-RFC2544 packet sizes would need to be approximated by those available in the table.
- o The Genome for very long sequences can become undecipherable by humans.

- o Whether h=MTU is useful/desirable is TBD.

- o Whether more tabulated packet sizes would be useful is TBD.

Some open issues with this format are:

1. Multiple Source-Destination Address Pairs: is the IMIX sequence applicable to each pair, across multiple pairs in sets, or across all pairs?

2. Multiple Tester Ports:is the IMIX sequence applicable to each port, across multiple ports in sets, or across all ports?

4. Reporting Long or Pseudo-Random Packet Sequences

When the IMIX-Genome cannot be used (when the sheer length of the sequence would make the genome unmanageable) or when the sequence is designed to vary within some proportional constraints, a table is necessary.

IP Length	Percentage of Total	Other Length(s)
64	23	82
128	67	146
1000	10	1018

Note that this approach also allows non-standard packet sizes, but trades the short genome specification and ability to specify the exact sequence for other flexibilities.

>>> Specification for psuedo-random size generation here? <<<

5. 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 [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.

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.

6. IANA Considerations

This memo makes no requests of IANA, and hopes that IANA will leave it alone as well.

7. Acknowledgements

8. References

8.1. Normative References

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

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

8.2. Informative References

[Agilent] http://www.ixiacom.com/pdfs/test_plans/agilent_journal_of_internet_test_methodologies.pdf, "The Journal of Internet Test Methodologies", 2007.

[IMIXonList] http://www.ietf.org/mail-archive/web/bmwg/current/msg00691.html, "Discussion on IMIX", 2003.

[IXIA] http://www.ixiacom.com/library/test_plans/display?skey=testing_pppox, "Library: Test Plans", 2010.

[Spirent] http://gospirent.com/whitepaper/IMIX%20Test%20Methodolgy%20Journal.pdf, "Test Methodology Journal: IMIX (Internet Mix) Journal", 2006.

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
URI: http://home.comcast.net/˜acmacm/

Benchmarking Methodology WG
Internet Draft
Intended status: Informational
Expires: April 2011

R.Papneja
Isocore

B.Parise
Cisco Systems

Susan Hares
Huawei

October 18, 2010

Basic BGP Convergence Benchmarking Methodology for Data Plane Convergence

draft-papneja-bgp-basic-dp-convergence-00.txt

Abstract

BGP is widely deployed and used by several service providers as the default Inter AS routing protocol. It is of utmost importance to ensure that when a BGP peer or a downstream link of a BGP peer fails, the alternate paths are rapidly used and routes via these alternate paths are installed. This document provides the basic BGP Benchmarking Methodology using existing BGP Convergence Terminology, RFC-4098.

Copyright Notice

Table of Contents

1. Introduction

This document defines the methodology for benchmarking data plane FIB convergence performance of BGP in router and switches for simple topologies of 3 or 4 nodes.

The methodology proposed in this document applies to both IPv4 and IPv6 and if a particular test is unique to one version, it is marked accordingly. For IPv6 benchmarking the device under test will require the support of Multi-Protocol BGP (MP-BGP) [RFC2858, RFC2545].

The scope of this companion document is limited to basic BGP protocol FIB convergence measurements. BGP extensions outside of carrying IPv6 in (MP-BGP) [RFC2858, RFC2545] are outside the scope of this document. Interaction with IGPs (IGP interworking) is outside the scope of this document.

1.1 Precise Benchmarking definition

Since benchmarking is science of precision, let us restate the purpose of this document in benchmarking terms. This document defines methodology to test

- data plane convergence on a single BGP device that supports the BGP [RFC4271] functionality;

  - in test topology of 3 or 4 nodes,

  - using Basic BGP.

Data plane convergence is defined as the completion of all FIB changes so that all forwarded traffic now takes the new proposed route. RFC 4098 defines the terms BGP device, FIB and the forwarded traffic. Data plane convergence is different than control plane convergence within a node.

Basic BGP is defined as RFC 4271 functional with Multi-Protocol BGP (MP-BGP) [RFC2858, RFC2545] for IPv6. The use of other extensions of BGP to support layer-2, layer-3 virtual private networks (VPN) are out of scope of this document.

The terminology used in this document is defined in [RFC4098]. One additional term is defined in this draft: data plane BGP convergence.

1.2 Purpose of BGP FIB (data plane) convergence

In the current Internet architecture the Inter-Autonomous System (inter-AS) transit is primarily available through BGP. To maintain a reliable connectivity within intra-domains or across inter-domains, fast recovery from failures remains most critical. To ensure minimal traffic losses, many service providers are requiring BGP implementations to converge the entire Internet routing table within sub-seconds at FIB level.

Furthermore, to compare these numbers amongst various devices, service providers are also looking at ways to standardize the convergence measurement methods. This document offers test methods for simple topologies. These simple tests will provide a quick high-level check, of the BGP data plane convergence across multiple implementations.

1.2 Control Plane Convergence

The convergence of BGP occurs at two levels: RIB and FIB convergence. RFC 4098 defines terms for BGP control plane convergence. Methodologies which test control plane convergence are out of scope for this draft.

1.3 Benchmarking Testing

In order to ensure that the results obtained in tests are repeatable, careful setup of initial conditions and exact steps are required.

This document proposes these initial conditions, test steps, and result checking. To ensure uniformity of the results all optional parameters SHOULD be disabled and all settings SHOULD be changed to default, these may include BGP timers as well.

2. Existing definitions and requirements

RFC 1242, "Benchmarking Terminology for Network Interconnect Devices" [RFC1242] and RFC 2285, "Benchmarking Terminology for LAN Switching Devices" [RFC2285] SHOULD be reviewed in conjunction with this document. WLAN-specific terms and definitions are also provided in Clauses 3 and 4 of the IEEE 802.11 standard [802.11]. Commonly used terms may also be found in RFC 1983 [RFC1983].

For the sake of clarity and continuity, this document adopts the general template for benchmarking terminology set out in Section 2 of RFC 1242. Definitions are organized in alphabetical order, and grouped into sections for ease of reference.

The following terms are assumed to be taken as defined in RFC 1242 [RFC1242]: Throughput, Latency, Constant Load, Frame Loss Rate, and Overhead Behavior. In addition, the following terms are taken as defined in RFC 2285 [RFC2285]: Forwarding Rates, Maximum Forwarding Rate, Loads, Device Under Test (DUT), and System Under Test (SUT).

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].

3. Test Topologies

This section describes simple test setups for use in BGP benchmarking tests measuring convergence of the FIB (data plane) after the BGP updates has been received.

These simple test nodes have 3 or 4 nodes with the following configuration:

1. Basic Test Setup
2. Three node setup for iBGP or eBGP convergence
3. Setup for eBGP multihop test scenario
4. Four node setup for iBGP or eBGP convergence

Individual tests refer to these topologies.

Figures 1-4 use the following conventions

AS-X: Autonomous System X

Loopback Int: Loopback interface on the BGP enabled device
R2: Helper router

3.1. General Reference Topology

Emulator acts as 1 or more BGP peers for different testcases.

```
 -----------                               -------------
|           |   traffic interfaces        |             |
|           |-----------------------1---- | tx          |
|           |-----------------------2---- | tr1         |
|           |-----------------------3-----| tr2         |
|    DUT    |    routing interfaces       | Emulator    |
|           |                             |             |
|      Drr1|---------------------------  |Err1         |
|           |      BGP Peering            |             |
|      Drr2|----------------------------  |Err2         |
|           |      BGP Peering            |             |
 -----------                                -------------
```

Figure 1 Basic Test Setup

```
    -------------         -------------           -------------
    |            |        |           |           |           |
    |            |        |           |           |           |
    |   HLP      |        |  DUT      |           | Emulator  |
    |  (AS-X)    |--------| (AS-Y)    |-----------|  (AS-Z)   |
    |            |        |           |           |           |
    |            |        |           |           |           |
    |            |        |           |           |           |
    -------------         -------------           -------------
```

Figure 2 Three Node Setup for eBGP and iBGP Convergence

```
    -------------         -------------           -------------
    |            |        |           |           |           |
    |            |        |           |           |           |
    |    R1      |        |  DUT      |           | Emulator  |
    |  (AS-X)    |--------| (AS-Y)    |-----------|  (AS-Z)   |
    |            |        |           |           |           |
    |            |        |           |           |           |
    |            |        |           |           |           |
    -------------         -------------           -------------
         |Loopback-Int           |Loopback-Int
         |                       |
         +                       +
```

Figure 3 BGP Convergence for eBGP Multihop Scenario

```
   ----------      ----------     ----------      ----------
    |         |     |        |     |        |     |         |
    |         |     |        |     |        |     |         |
    |    R1   |     |  DUT2  |     |  DUT1  |     |Emulator |
    | (AS-X)  |-----| (AS-X) |-----| (AS-Y) |-----| (AS-Z)  |
    |         |     |        |     |        |     |         |
    |         |     |        |     |        |     |         |
    |         |     |        |     |        |     |         |
    ----------      ----------     ----------      ----------
```

Figure 4 Four Node Setup for EBGP and IBGP Convergence

4. Test Considerations

The test cases for measuring convergence for iBGP and eBGP are different. Both iBGP and eBGP use different mechanisms to advertise, install and learn the routes. Typically, an iBGP route on the DUT is installed and exported only when the next-hop is reachable. For eBGP the route is installed on the DUT with the remote interface address as the next-hop with the exception of the multihop case.

4.1. Number of Peers

Number of Peers is defined as the number of BGP neighbors or sessions the DUT has at the beginning of the test.
The peers are established before the tests begin.

The relationship could be either, iBGP or eBGP peering depending upon the test case requirement.

The DUT establishes one or more BGP sessions with one more emulated routers or helper nodes. Additional peers can be added based on the testing requirements. The number of peers enabled during the testing should be well documented in the report matrix.

4.2. Number of Routes per Peer

It Number of Routes per Peer is defined as the number of routes advertized or learnt by the DUT per session or through neighbor relationship with an emulator or helper node. The tester, emulating as neighbor MUST advertise at least one route per peer.

Each test must run must identify the route stream in terms of route packing, route mixture, and number of routes. This route stream must be well documented in the reporting stream. RFC 4098 defines these terms.

It is RECOMMENDED that the user may consider advertizing the entire current Internet routing table per peering session using an Internet route mixture with unique or non-unique routes.
If multiple peers are used, it is important to precisely document the timing sequence between the peer sending routes (as defined in RFC 4098).

4.3. Policy Processing/Reconfiguration

The DUT MUST run one baseline test where policy is Minimum policy as defined in RFC 4098. Additional runs may be done with policy set-up before the tests begin. Exact policy settings should be documented as part of the test.

4.4. Configured Parameters (Timers, etc..)

There are configured parameters and timers that may impact the measured BGP convergence times.

The benchmark metrics MAY be measured at any fixed values for these configured parameters.

It is RECOMMENDED these configure parameters have two settings: a) basic-test, and b)values as expected in the operational network. All optional BGP settings MUST be kept consistent across iterations of any specific tests

Examples of the configured parameters that may impact measured BGP convergence time include, but are not limited to:

1. Interface failure detection timer
2. BGP Keepalive timer
3. BGP Holdtime
4. BGP update delay timer
5. ConnectRetry timer
6. TCP Segment Size
7. Minimum Route Advertisement Interval (MRAI)
8. MinASOriginationInterval (MAOI)
9. Route Flap Dampening parameters
10. TCP MD5

The basic-test settings for the parameters should be:

1. Interface failure detection timer (0 ms)
2. BGP Keepalive timer (1 min)
3. BGP Holdtime (3 min)
4. BGP update delay timer (0 s)
5. ConnectRetry timer (1 s)
6. TCP Segment Size (4096)
7. Minimum Route Advertisement Interval (MRAI)(0 s)
8. MinASOriginationInterval (MAOI) (0 s)
9. Route Flap Dampening parameters (off)
10. TCP MD5 (off)

4.5. Interface Types

The type of media dictate which test cases may be executed, each interface type has unique mechanism for detecting link failures and the speed at which that mechanism operates will influence the measurement results. All interfaces MUST be of the same media and throughput for each test case.

4.6. Measurement Accuracy

Since observed packet loss is used to measure the route convergence time, the time between two successive packets offered to each individual route is the highest possible accuracy of any packet-loss based measurement. When packet jitter is much less than the convergence time, it is a negligible source of error and hence it will be treated as within tolerance.

An exterior measurement on the input media (such Ethernet)is defined by this specification.

4.7. Measurement Statistics

The benchmark measurements may vary for each trial, due to the statistical nature of timer expirations, CPU scheduling, etc. It is recommended to repeat the test multiple times. Evaluation of the test data must be done with an understanding of generally accepted testing practices regarding repeatability, variance and statistical significance of a small number of trials.

For any repeated tests that are averaged to remove variance, all parameters MUST remain the same.

4.8. Authentication

Authentication in BGP is done using the TCP MD5 Signature Option [RFC2385]. The processing of the MD5 hash, particularly in devices with a large number of BGP peers and a large amount of update traffic, can have an impact on the control plane of the device. If authentication is enabled, it SHOULD be documented correctly in the reporting format

4.9. Convergence Events

Convergence events or triggers are defined as abnormal occurrences in the network, which initiate route flapping in the network, and hence forces the re-convergence of a steady state network. In a real network, a series of convergence events may cause convergence latency operators desire to test.

These convergence events must be defined in terms of the sequences defined in RFC 4098. This basic document begins all tests with a router initial set-up. Additional documents will define BGP data plane convergence based on peer initialization.

The convergence events may or may not be tied to the actual failure A Soft Reset (RFC 4098) does not clear the RIB or FIB tables. A Hard reset clears the BGP peer sessions, the RIB tables, and FIB tables.

4.10. High Availability

Due to the different Non-Stop-Routing (sometimes referred to High-Availability) solutions available from different vendors, it is RECOMMENDED that any redundancy available in the routing processors should be disabled during the convergence measurements.

5. Test Cases

All tests defined under this section assume the following:

BGP peers should be brought to BGP Peer established state.

a. Furthermore the traffic generation and routing should be verified in the topology

5.1. Basic Convergence Tests

These test cases measure characteristics of a BGP implementation in non-failure scenarios like:

a. RIB-IN Convergence

b. RIB-OUT Convergence

c. eBGP Convergence

d. iBGP Convergence

5.1.1. RIB-IN Convergence

Objective:

This test measures the convergence time taken to receive and install a route in RIB using BGP

Reference Test Setup:

This test uses the setup as shown in figure 1

Procedure:

a. All variables affecting Convergence should be set to a basic test state (as defined in section 4-4).
b. Establish BGP adjacency between DUT and peer x of Emulator;
c. To ensure adjacency establishment, wait for 3 KeepAlives from the DUT or a configurable delay before proceeding with the rest of the test;
d. Start the traffic from the Emulator peer-x towards the DUT targeted at a routes specified in route mixture (ex. route A) Initially no traffic SHOULD be observed on the egress interface as the route A is not installed in the forwarding database of the DUT.
e. Advertise route A from the Peer-x to the DUT and record the time;

   This is Tup(EMx,Rt-A). (nick-name XMT-Rt-time)

f. Record the time when the route-A from Peer-x is received at the DUT.

   This Tup(DUT,Rt-A). It is nick named is RCV-Rt-time

g. Record the time when the traffic targeted towards route A is received by Emulator on appropriate traffic egress interface.
   If 3rd party route (traffic-egress 2), or BGP peer route interfaces.

   This is TR(TDx,Rt-A). This is "nick-named" DUT-XMT-Data-Time.

h. The difference between the Tup(TDx,RT-A) and traffic received time (TR (TDr, Rt-A) is the FIB Convergence Time for route-A in the route mixture.

   A full convergence for the route update is the measurement between the 1st route (Route-A) and the last route (Rt-last)

   Route update convergence is
   TR(TDr, RT-last)- Tup(DUT, Rt-A) or

   (DUT-XMT-Data-Time - RCV-Rt-Time)(rt-A)

Note: It is recommended that a single test with the same route mixture be repeated several times. A report should provide the Stand deviation of all tests and the average.

Running tests with a varying number of routes and route mixtures is important to get a full characterization of a single peer.

5.1.2. RIB-OUT Convergence

Objective:

This test measures the convergence time taken by an implementation to receive, install and advertise a route using BGP

Reference Test Setup:

This test uses the setup as shown in figure 2

Procedure:

a. The Helper node (HLP) run same version of BGP as DUT;
b. All devices MUST be synchronized using NTP or some local reference clock;
c. All configuration variables for HLP, DUT, and Emulator SHOULD be set to the same values. These values MAY be basic-test or a unique set completely described in the test set-up.
d. Establish BGP adjacency between DUT and Emulator
e. Establish BGP adjacency between DUT and Helper Node
f. To ensure adjacency establishment, wait for 3 KeepAlives from the DUT or a configurable delay before proceeding with the rest of the test
g. Start the traffic from the Emulator towards the Helper Node targeted at a specific route say route A.
   Initially no traffic SHOULD be observed on the egress interface as the route-A is not installed in the forwarding database of the DUT
h. Advertise routeA from the Emulator to the DUT and note the time.
   This is Tup(EMx, Route-A). (nick-name EM-XMT-Rt-Time)

i. Record when Route-A is received by DUT.

This is Tup(DUTr, Route-A). (nick-name DUT-RCV-Rt-Time)

j. Record the time when the ROUTE forward by DUT toward the Helper node.

This is Tup(DUTx, Rt-A). (nick-name DUT-XMT-Rt-Time).

k. Record the time when the traffic targeted towards route-A is received on the Route Egress Interface toward peer-X. This is TR(EMr, Route-A). (nick-name DUT-XMT-Data Time).

FIB convergence = (DUT-RCV-Rt-Time - DUT-XMT-Data-Time).

RIB convergence = (DUT-RCV-Rt-Time - DUT-XMT-Rt-Time).

Convergence for a route stream is characterized by

a) Individual route convergence for FIB, RIB
b) All route convergence of

FIB-convergence =DUT-RCV-Rt-Time(A)-DUT-XMT-Data-Time(last)

RIB-convergence =DUT-RCV-Rt-Time(A)-DUT-XMT-Rt-Time(last)

5.1.3. eBGP Convergence

Objective:

This test measures the convergence time taken by an implementation to receive, install and advertise a route in an eBGP Scenario

Reference Test Setup:

This test uses the setup as shown in figure 2, and the scenarios described in RIB-IN and RIB-OUT are applicable to this test case.

5.1.4. iBGP Convergence

Objective:

This test measures the convergence time taken by an implementation to receive, install and advertise a route in an iBGP Scenario

Reference Test Setup:

This test uses the setup as shown in figure 2, and the test scenarios listed in RIB-IN and RIB-OUT are applicable to this test case.

5.1.5. eBGP Multihop Convergence

Objective

This test measures the convergence time taken by an implementation to receive, install and advertise a route in an eBGP Multihop Scenario

Reference Test Setup:

This test uses the setup as shown in figure 3. Two DUTs are used along with a helper node.

Procedure:

a. The DUT2 is the same model as DUT and runs the same BGP implementation as DUT.
b. All devices to be synchronized using NTP
c. All variables affecting Convergence like authentication, policies, timers should be set to basic-settings.
d. All 3 devices, DUT, Emulator and Helper Node are configured as different Autonomous Systems
e. Loopback Interfaces configured on DUT and Helper Node and connectivity is established between them using any config options available on the DUT
f. Establish BGP adjacency between DUT1 and Emulator
g. Establish BGP adjacency between DUT2 and Helper Node
h. Establish BGP adjacency between DUT 1 and DUT 2
i. To ensure adjacency establishment, wait for 3 KeepAlives from the DUT1 and DUT2 or a configurable delay before proceeding with the rest of the test
j. Start the traffic from the Emulator towards the Helper Node targeted at a specific route say routeA.

k. Initially no traffic SHOULD be observed on the egress interface as the routeA is not installed in the forwarding database of the DUT
l. Advertise routeA from the Emulator to the DUT and note the time. (Tup(EMx,RouteA) - this is nicknamed (Route-Rec-time).

m. Record the time when the traffic targeted towards routeA is received from Egress Interface of DUT on emulator.

   This is TR(EMr,DUT), nicknamed (Data Receive time)

n. The following equation represents the FIB Convergence multi-node

   eBGP Multihop Convergence Time =
   (Rt-RecTime - Data-RcvTime).

Note: It is recommended that the test be repeated with varying number of routes and route mixtures. With each set route mixture, the test should be repeated multiple times. The results should record average, mean, Standard Deviation.

5.2. BGP Failure/Convergence Events

5.2.1. Physical Link Failure on DUT End

Objective:

This test measures the route convergence time due to local link failure event at DUT’s Local Interface

Reference Test Setup:

This test uses the setup as shown in figure 1. Shutdown event is defined as an administrative shutdown event on the DUT.

Procedure:

a. All variables affecting Convergence like authentication, policies, timers should be set to basic-test policy.

b. Establish 2 BGP adjacencies from DUT to Emulator, one over the peer interface and the other using a second peer interface.

c. Advertise the same route, route A over both the adjacencies and (Tx1)Interface to be the preferred next hop.

d. To ensure adjacency establishment, wait for 3 KeepAlives from the DUT or a configurable delay before proceeding with the rest of the test.

e. Start the traffic from the Emulator towards the DUT targeted at a specific route say route A. Initially traffic would be observed on the best egress route (Err1) instead of Trr2

f. Trigger the shutdown event of Best Egress Interface on DUT (Drr1).

g. Measure the Convergence Time for the event to be detected and traffic to be forwarded to Next-Best Egress Interface (rr2).

   Time = Data-detect(rr2) - Shutdown time.

h. Stop the offered load and wait for the queues to drain and Restart

i. Bring up the link on DUT Best Egress Interface

j. Measure the convergence time taken for the traffic to be rerouted from (rr2) to Best Interface (rr1)

   Time = Data-Detect(rr1) - Shutdown-time.

k. It is recommended that the test be repeated with varying number of routes and route mixtures or with number of routes & route mixtures closer to what is deployed in operational networks

5.2.2. Physical Link Failure on Remote/Emulator End

Objective:

This test measures the route convergence time due to local link failure event at Tester’s Local Interface

Reference Test Setup:

This test uses the setup as shown in figure 1. Shutdown event is defined as shutdown of the local interface of Tester via logical shutdown event. The procedure used in 5.2.1 is used for the termination.

5.2.3. ECMP Link Failure on DUT End

Objective:

This test measures the route convergence time due to local link failure event at ECMP Member. The FIB configuration and BGP is set to allow two ECMP routes to be installed. However, policy directs the routes to be sent only over one of the paths.

Reference Test Setup:

This test uses the setup as shown in figure 1, and the procedure uses 5.2.1.

5.3. BGP Adjacency Failure (Non-Physical Link Failure) on Emulator

Objective:

This test measures the route convergence time due to BGP Adjacency Failure on Emulator

Reference Test Setup:

This test uses the setup as shown in figure 1

Procedure:

a. All variables affecting Convergence like authentication, policies, timers should be basic-policy set.

b. Establish 2 BGP adjacencies from DUT to Emulator, one over the Best Egress Interface and the other using the Next-Best Egress Interface
c. Advertise the same route, routeA over both the adjacencies and make Best Egress Interface to be the preferred next hop
d. To ensure adjacency establishment, wait for 3 KeepAlives from the DUT or a configurable delay before proceeding with the rest of the test
e. Start the traffic from the Emulator towards the DUT targeted at a specific route say routeA. Initially traffic would be observed on the Best Egress interface
f. Remove BGP adjacency via a software adjacency down on the Emulator on the Best Egress Interface

   Time = BGPadj-down-time - nicknamed BGPpeer-down.

g. Measure the Convergence Time for the event to be detected and traffic to be forwarded to Next-Best Egress Interface

   This time is Tr-rr2 nicknamed - TR2-traffic-on

   Convergence = TR2-traffic-on - BGPpeer-down

h. Stop the offered load and wait for the queues to drain and Restart
i. Bring up BGP adjacency on the Emulator over the Best Egress Interface
   Time = BGP-adj-up - nicknamed BGPpeer-up

j. Measure the convergence time taken for the traffic to be rerouted to Best Interface

   Time = Tr-rr1 is nicknamed TR1-traffic-on.

5.4. BGP Hard Reset Test cases

5.4.1. BGP Non-Recovering Hard Reset Event on DUT

Objective:

This test measures the route convergence time due to Hard Reset on the DUT

Reference Test Setup:

This test uses the setup as shown in figure 1

Procedure:

a. The requirement for this test case is that the Hard Reset Event should be non-recovering and should affect only the adjacency between DUT and Emulator on the Best Egress Interface
b. All variables affecting SHOULD be set to basic-test values
c. Establish 2 BGP adjacencies from DUT to Emulator, one over the Best Egress Interface and the other using the Next-Best Egress Interface
d. Advertise the same route, routeA over both the adjacencies and make Best Egress Interface to be the preferred next hop
e. To ensure adjacency establishment, wait for 3 KeepAlives from the DUT or a configurable delay before proceeding with the rest of the test
f. Start the traffic from the Emulator towards the DUT targeted at a specific route say routeA. Initially traffic would be observed on the Best Egress interface
g. Trigger the Hard Reset event of Best Egress Interface on DUT
h. Measure the Convergence Time for the event to be detected and traffic to be forwarded to Next-Best Egress Interface

   Time of convergence = time-traffic flow - time-reset.

i. Stop the offered load and wait for the queues to drain and Restart
j. It is recommended that the test be repeated with varying number of routes and route mixtures or with number of routes & route mixtures closer to what is deployed in operational networks
k. When varying number of routes are used, convergence Time is measured using the Loss Derived method [IGP-Data]

l. Convergence Time in this scenario is influenced by Failure detection time on Tester, BGP Keep Alive Time and routing, forwarding table update time

5.5. BGP Soft Reset

Objective:

This test measures the route convergence time taken by an implementation to service a BGP Route Refresh message and advertise a route

Reference Test Setup:

This test uses the setup as shown in figure 2

Procedure:

a. The BGP implementation on DUT & Helper Node needs to support BGP Route Refresh Capability [RFC 2918]
b. All devices to be synchronized using NTP
c. All variables affecting Convergence like authentication, policies, timers should be set to basic-test defaults.
d. DUT and Helper Node are configured in the same Autonomous System whereas Emulator is configured under a different Autonomous System
e. Establish BGP adjacency between DUT and Emulator
f. Establish BGP adjacency between DUT and Helper Node
g. To ensure adjacency establishment, wait for 3 KeepAlives from the DUT or a configurable delay before proceeding with the rest of the test
h. Configure a policy under BGP on Helper Node to deny routes received from DUT
i. Advertise routeA from the Emulator to the DUT
j. The DUT will try to advertise the route to Helper Node will be denied
k. Wait for 3 KeepAlives
l. Start the traffic from the Emulator towards the Helper Node targeted at a specific route say routeA. Initially no traffic

would be observed on the Egress interface, as routeA is not present

m. Remove the policy on Helper Node and issue a Route Refresh request towards DUT. Note the timestamp of this event. This is the RefreshTime
n. Record the time when the traffic targeted towards routeA is received on the Egress Interface. This is RecTime
o. The following equation represents the Route Refresh Convergence Time per route
    i. Route Refresh Convergence Time = (RecTime - RefreshTime)

5.6. BGP Route Withdrawal Convergence Time

Objective:

This test measures the route convergence time taken by an implementation to service a BGP Withdraw message and advertise the withdraw

Reference Test Setup:

This test uses the setup as shown in figure 2

Procedure:

a. This test consists of 2 steps to determine the Total Withdraw Processing Time
b. Step 1:
    i. All devices to be synchronized using NTP
    ii. All variables should be set to basic-test parameters.
    iii. DUT and Helper Node are configured in the same Autonomous System whereas Emulator is configured under a different Autonomous System
    iv. Establish BGP adjacency between DUT and Emulator
    v. To ensure adjacency establishment, wait for 3 KeepAlives from the DUT or a configurable delay before proceeding with the rest of the test
    vi. Start the traffic from the Emulator towards the DUT targeted at a specific route say routeA. Initially no

traffic would be observed on the Egress interface as the routeA is not present on DUT.

vii. Advertise routeA from the Emulator to the DUT

viii. The traffic targeted towards routeA is received on the Egress Interface

ix. Now the Tester sends request to withdraw routeA to DUT. TRx(Awith) nicknamed WdrawTime1

x. Record the time when no traffic is observed on the Egress Interface.

This is the RouteRemoveTime1(A)

WdrawConvTime1 = RouteRemoveTime1(A)

xi. The difference between the RouteRemoveTime1 and WdrawTime1 is the WdrawConvTime1

c. Step 2:

i. Continuing from Step 1, re-advertise routeA back to DUT from Tester

ii. The DUT will try to advertise the routeA to Helper Node (assumption there exists a session between DUT and helper node)

iii. Start the traffic from the Emulator towards the Helper Node targeted at a specific route say routeA. Traffic would be observed on the Egress interface after routeA is received by the Helper Node

WATime=time traffic first flows

iv. Now the Tester sends a request to withdraw routeA to DUT. This is the WdrawTime2

WAWtime-TRx(RouteA) = is nicknamed WdrawTime2

v. DUT processes the withdraw and sends it to Helper Node

vi. Record the time when no traffic is observed on the Egress Interface of Helper Node. This is the

TR-WAW(DUT,RouteA) = RouteRemoveTime2

vii. Total withdraw processing time is

TotalWdrawTime = ((RouteRemoveTime2 - WdrawTime2) - WdrawConvTime1)

5.7. BGP Path Attribute Change Convergence Time

Objective:

This test measures the route convergence time taken by an implementation to service a BGP Path Attribute Change

Reference Test Setup:

This test uses the setup as shown in figure 1

Procedure:

a. This test only applies to Well-Known Mandatory Attributes like Origin, AS Path, Next Hop
b. In each iteration of test only one of these mandatory attributes need to be varied whereas the others remain the same
c. All devices to be synchronized using NTP
d. All variables should be set to basic-test parameters
e. Advertise the route, routeA over the Best Egress Interface only, making it the preferred next hop
f. To ensure adjacency establishment, wait for 3 KeepAlives from the DUT or a configurable delay before proceeding with the rest of the test
g. Start the traffic from the Emulator towards the DUT targeted at the specific route say routeA. Initially traffic would be observed on the Best Egress interface
h. Now advertise the same route routeA on the Next-Best Egress Interface but by varying one of the well-known mandatory attributes to have a preferred value over that interface. The

other values need to be same as what was advertised on the Best-Egress adjacency.

TRx(Path-Change) = Path Change Event Time

i. Measure the Convergence Time for the event to be detected and traffic to be forwarded to Next-Best Egress Interface

DUT(Path-Change, RouteA) = Path-switch time

Convergence = Path-switch time - Path Change Event Time.

j. Stop the offered load and wait for the queues to drain and Restart

5.8. BGP Graceful Restart Convergence Time

Objective:

This test measures the route convergence time taken by an implementation during a Graceful Restart Event

Reference Test Setup:

This test uses the setup as shown in figure 4

Procedure:

a. It measures the time taken by an implementation to service a BGP Graceful Restart Event and advertise a route
b. The Helper Nodes are the same model as DUT and run the same BGP implementation as DUT
c. The BGP implementation on DUT & Helper Node needs to support BGP Graceful Restart Mechanism [RFC4724]
d. All devices to be synchronized using NTP
e. All variables are set to basic-test values.
f. DUT and Helper Node-1 are configured in the same Autonomous System whereas Emulator and Helper Node-2 are configured under different Autonomous Systems
g. Establish BGP adjacency between DUT and Helper Nodes

h. Establish BGP adjacency between Helper Node-2 and Emulator
i. To ensure adjacency establishment, wait for 3 KeepAlives from the DUT or a configurable delay before proceeding with the rest of the test
j. Configure a policy under BGP on Helper Node-1 to deny routes received from DUT
k. Advertise routeA from the Emulator to Helper Node-2
l. Helper Node-2 advertises the route to DUT and DUT will try to advertise the route to Helper Node-1 which will be denied
m. Wait for 3 KeepAlives
n. Start the traffic from the Emulator towards the Helper Node-1 targeted at the specific route say routeA. Initially no traffic would be observed on the Egress interface as the routeA is not present
o. Perform a Graceful Restart Trigger Event on DUT and note the time. This is the GREventTime
p. Remove the policy on Helper Node-1
q. Record the time when the traffic targeted towards routeA is received on the Egress Interface.

   TRr(DUT, routeA). This is nicknamed RecTime.

r. The following equation represents the Graceful Restart Convergence Time
   i. Graceful Restart Convergence Time = ((GREventTime - RecTime) - RIB-IN)

s. It is assumed in this test case that after a Switchover is triggered on the DUT, it will not have any cycles to process BGP Refresh messages.
   The reason for this assumption is that there is a narrow window of time where after switchover when we remove the policy from Helper Node -1, implementations might generate Route-Refresh automatically and this request might be serviced before the DUT actually switches over and reestablishes BGP adjacencies with the peers

6. Reporting Format

For each test case, it is recommended that the reporting tables below are completed and all time values SHOULD be reported with resolution as specified in [RFC 4098].

Parameter	Units
Test case	Test case number
Test topology	1,2,3 or 4
Parallel links	Number of parallel links
Interface type	GigE, POS, ATM, other
Convergence Event	Hard reset, Soft reset, link failure, or other defined
eBGP sessions	Number of eBGP sessions
iBGP sessions	Number of iBGP sessions
eBGP neighbor	Number of eBGP neighbors
iBGP neighbor	Number of iBGP neighbors
Routes per peer	Number of routes
Total unique routes	Number of routes
Total non-unique routes	Number of routes
IGP configured	ISIS, OSPF, static, or other
Route Mixture	Description of Route mixture
Route Packing	Number of routes in an update
Policy configured	Yes, No
Packet size offered to the DUT	Bytes
Offered load	Packets per second
Packet sampling interval on tester	Seconds
Forwarding delay threshold	Seconds
Timer value configured on DUT	
Interface failure indication delay	Seconds
Hold timer	Seconds
MinRouteAdvertisementInterval (MRAI)	Seconds
MinASOriginationInterval (MAOI)	Seconds
Keepalive	Seconds
ConnectRetry	Seconds
TCP Parameters for DUT and	

tester

MSS	Bytes
Slow start threshold	Bytes
Maximum window size	Bytes

Test Details:

a. If the Offered Load matches a subset of routes, describe how this subset is selected.
b. Describe how the Convergence Event is applied; does it cause instantaneous traffic loss or not.
c. If there is any policy configured, describe the configured policy.

Complete the table below for the initial Convergence Event and the reversion Convergence Event.

Parameter	Unit
Conversion Event	Initial or reversion
Traffic Forwarding Metrics	
Total number of packets offered to DUT	Number of packets
Total number of packets forwarded by DUT	Number of packets
Connectivity Packet Loss	Number of packets
Convergence Packet Loss	Number of packets
Out-of-order packets	Number of packets
Duplicate packets	Number of packets
Convergence Benchmarks	
Rate-derived Method[IGP-Data]:	
First route convergence time	Seconds
Full convergence time	Seconds
Loss-derived Method [IGP-Data]:	
Loss-derived convergence time	Seconds
Route-Specific Loss-Derived Method:	

```
        Minimum R-S convergence    Seconds
         time
         Maximum R-S convergence    Seconds
         time
         Median R-S convergence     Seconds
         time
         Average R-S convergence    Seconds
         time

    Loss of Connectivity Benchmarks
       Loss-derived Method:
         Loss-derived loss of       Seconds
         connectivity period
       Route-Specific loss-derived
       Method:
         Minimum LoC period [n]     Array of seconds
         Minimum Route LoC period   Seconds
         Maximum Route LoC period   Seconds
         Median Route LoC period    Seconds
         Average Route LoC period   Seconds
```

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 Considerations

This document requires no IANA considerations.

9. References

9.1. Normative References

[RFC1771] Rekhter, Y. and Li, T., "A Border Gateway Protocol 4(BGP-4)", RFC 4271, March 1995.

[RFC4098] Berkowitz, H. et al., "Terminology for benchmarking BGP device convergence in control plane", RFC4098, June 2005

[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.

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

[IGP-Data] Poretsky, S. et al., "Terminology for benchmarking Link-state IGP data plane convergence," draft-ietf-bmwg-igp-dataplane-conv-term-21, May 2010

9.2. Informative References

[RFC2858] Bates, T. et al., "Multiprotocol extensions for BGP-4," RFC 2858, June 2000

[RFC2545] Marques, P. and F. Dupont, "Use of BGP-4 multiprotocol extensions for IPv6 Inter-Domain Routing," RFC2545, March 1999

[RFC3107] Rekhter, Y. and E. Rosen, "Carrying label information in BGP-4," RFC 3107, May 2001

[RFC2385] Heffernan, A., "Protection of BGP Sessions via the TCP MD5 signature option," RFC2385, August 1998

[RFC2918] Chen, E., "Route Refresh capability for BGP-4," RFC 2918, September 2000

[RFC4724] Sangli, S. et al., "Graceful restart mechanism for BGP," RFC 4724, Jan 2007

[RFC1983] Malkin, G.,"Internet Users' Glossary", RFC 1983, August 1996

Authors Addresses

Rajiv Papneja
Isocore
12359 Sunrise Valley Dr. STE100
Reston, VA 20191
Phone: +1.703.860.9273
Email: rpapneja@isocore.com

Bhavani Parise
Cisco Systems
170 West Tasman Drive
San Jose, CA 95134
Phone: +408-853-6346
Email: bhavani@cisco.com

Susan Hares
Huawei Technologies (USA)
2330 Central Expressway
Santa Clara, CA 95050
Phone: +408-330-4581
Cell: +1-734-604-0332
Email shares@huawei.com

Eric Brendel
Independent Consultant
154 3rd St,
Fair Haven, NJ 07704
Phone:+1.732.895.1504
Email: brendel@pektel.com

Mohan Nanduri
Microsoft
12012 Sunset Hills Rd.
Reston, VA 20190
Phone: +1.703.627.6455
Email: mnanduri@microsoft.com

Jay Karthik
Cisco Systems
170 West Tasman Drive
San Jose, CA 95134
Phone: +1.978.319.0527
Email: jkarthik@cisco.com

Network Working Group
Internet-Draft
Intended status: Informational
Expires: April 22, 2011

T. Player
Spirent Communications
D. Newman
Network Test

October 19, 2010

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

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