

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

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

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

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

This Internet-Draft will expire on May 9, 2013.

Copyright Notice

Copyright (c) 2012 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

This document may contain material from IETF Documents or IETF Contributions published or made publicly available before November 10, 2008. The person(s) controlling the copyright in some of this material may not have granted the IETF Trust the right to allow modifications of such material outside the IETF Standards Process. Without obtaining an adequate license from the person(s) controlling the copyright in such materials, this document may not be modified outside the IETF Standards Process, and derivative works of it may not be created outside the IETF Standards Process, except to format it for publication as an RFC or to translate it into languages other than English.

Table of Contents

1. Introduction	5
2. Document Scope	6
3. Existing Definitions and Requirements	6
4. General Reference Topology	7
5. Test Considerations	8
5.1. Failover Events [RFC 6414]	8
5.2. Failure Detection [RFC 6414]	9
5.3. Use of Data Traffic for MPLS Protection benchmarking	10
5.4. LSP and Route Scaling	10
5.5. Selection of IGP	10
5.6. Restoration and Reversion [RFC 6414]	10
5.7. Offered Load	11
5.8. Tester Capabilities	11
5.9. Failover Time Measurement Methods	12
6. Reference Test Setup	12
6.1. Link Protection	13
6.1.1. Link Protection - 1 hop primary (from PLR) and 1 hop backup TE tunnels	13
6.1.2. Link Protection - 1 hop primary (from PLR) and 2 hop backup TE tunnels	14
6.1.3. Link Protection - 2+ hop (from PLR) primary and 1 hop backup TE tunnels	14
6.1.4. Link Protection - 2+ hop (from PLR) primary and 2 hop backup TE tunnels	15
6.2. Node Protection	16
6.2.1. Node Protection - 2 hop primary (from PLR) and 1 hop backup TE tunnels	16
6.2.2. Node Protection - 2 hop primary (from PLR) and 2 hop backup TE tunnels	17
6.2.3. Node Protection - 3+ hop primary (from PLR) and 1 hop backup TE tunnels	18
6.2.4. Node Protection - 3+ hop primary (from PLR) and 2 hop backup TE tunnels	19
7. Test Methodology	20
7.1. MPLS FRR Forwarding Performance	20
7.1.1. Headend PLR Forwarding Performance	20
7.1.2. Mid-Point PLR Forwarding Performance	21
7.2. Headend PLR with Link Failure	23
7.3. Mid-Point PLR with Link Failure	24
7.4. Headend PLR with Node Failure	26
7.5. Mid-Point PLR with Node Failure	27
8. Reporting Format	28
9. Security Considerations	30
10. IANA Considerations	30
11. Acknowledgements	30
12. References	30

12.1. Informative References	30
12.2. Normative References	30
Appendix A. Fast Reroute Scalability Table	30
Appendix B. Abbreviations	33
Authors' Addresses	34

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 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.

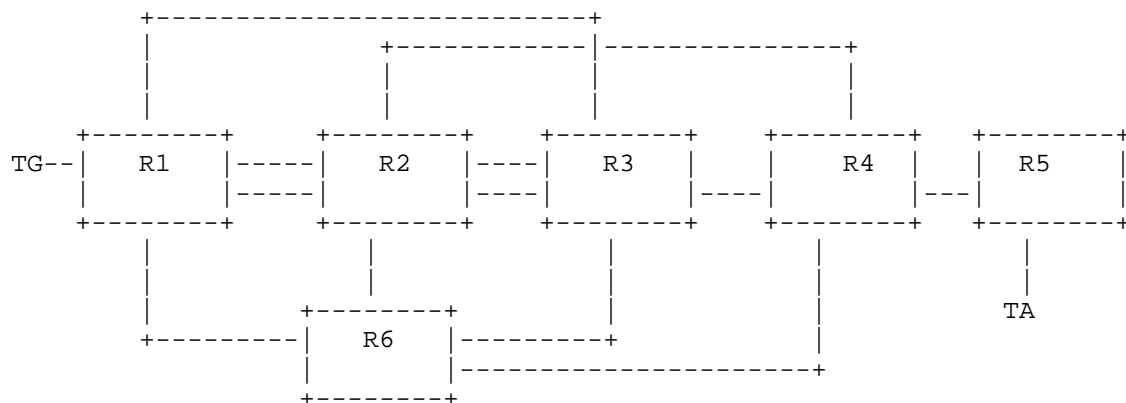


Fig. 1 Fast Reroute Topology

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 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

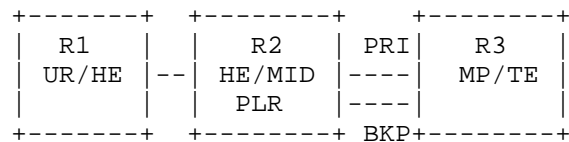


Figure 2.

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

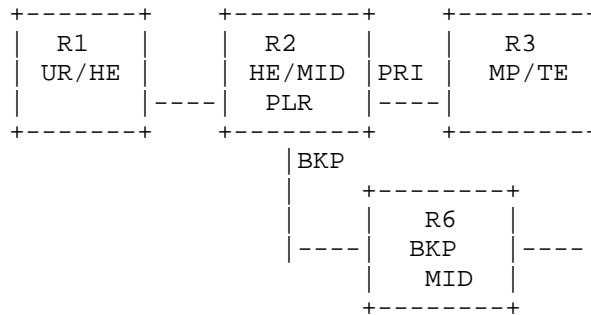


Figure 3.

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

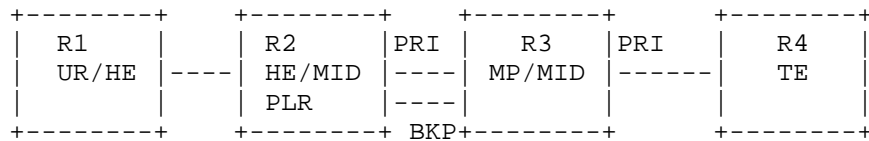


Figure 4.

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

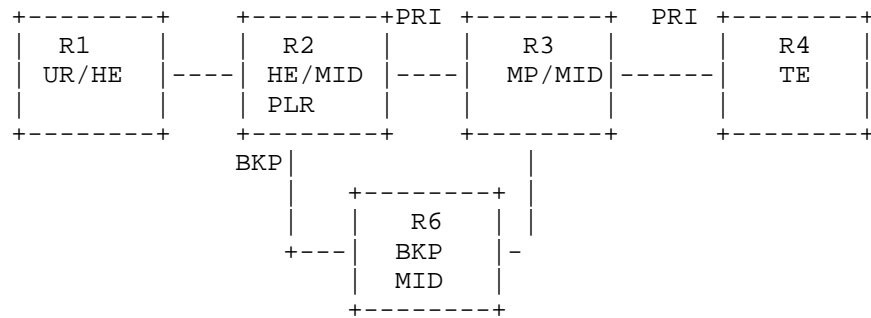


Figure 5.

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

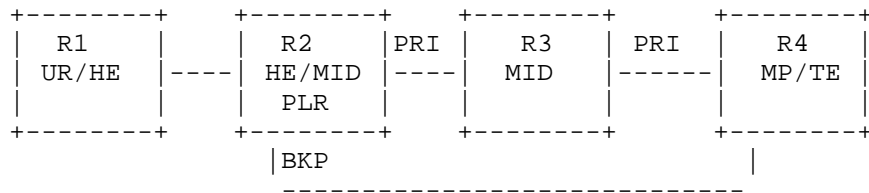


Figure 6.

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

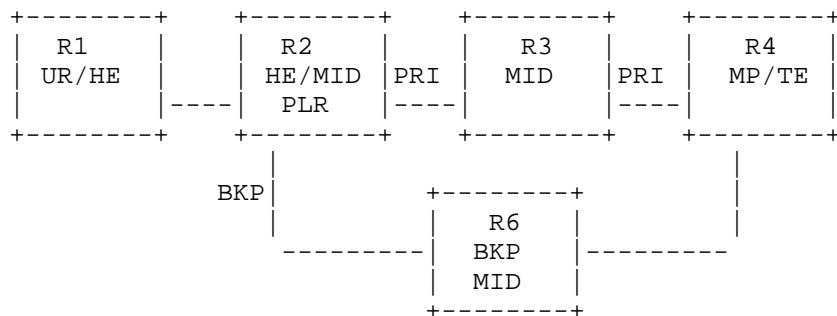


Figure 7.

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 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

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

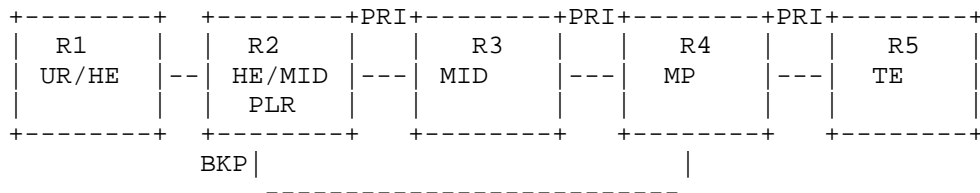


Figure 8.

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

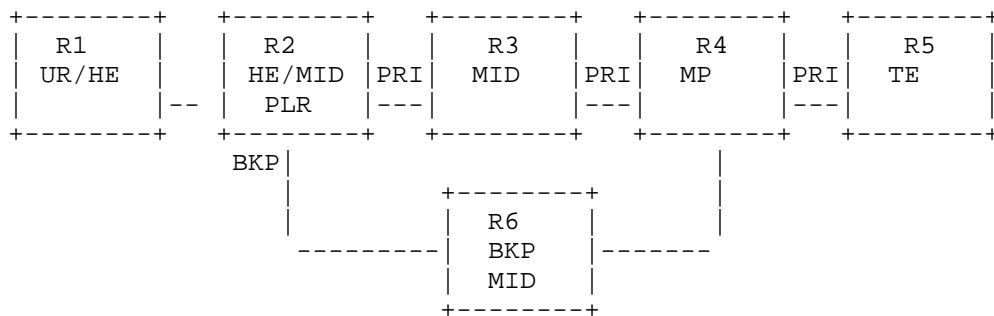


Figure 9.

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

Network Working Group
Internet Draft
Expires: Jan 2011
Intended Status: Informational

S. Poretsky
Allot Communications
Rajiv Papneja
Isocore
J. Karthik
S. Vapiwala
Cisco Systems
July 2010

Benchmarking Terminology
for Protection Performance
<draft-ietf-bmwg-protection-term-09.txt >

Status of this Memo

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

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

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

The list of current Internet-Drafts can be accessed at
<http://www.ietf.org/ietf/lid-abstracts.txt>.
The list of Internet-Draft Shadow Directories can be accessed at
<http://www.ietf.org/shadow.html>.

This Internet-Draft will expire on 7 Jan, 2011.

Copyright Notice

Copyright (c) 2010 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

This document may contain material from IETF Documents or IETF Contributions published or made publicly available before November 10, 2008. The person(s) controlling the copyright in some of this material may not have granted the IETF Trust the right to allow modifications of such material outside the IETF Standards Process. Without obtaining an adequate license from the person(s) controlling the copyright in such materials, this document may not be modified outside the IETF Standards Process, and derivative works of it may not be created outside the IETF Standards Process, except to format it for publication as an RFC or to translate it into languages other than English.

Abstract

This document provides common terminology and metrics for benchmarking the performance of sub-IP layer protection mechanisms. The performance benchmarks are measured at the IP-Layer with protection may be provided at the Sub-IP layer. The benchmarks and terminology can be applied in methodology documents for different sub-IP layer protection mechanisms such as Automatic Protection Switching (APS), Virtual Router Redundancy Protocol (VRRP), Stateful High Availability (HA), and Multi-Protocol Label Switching Fast Reroute (MPLS-FRR).

Table of Contents

1. Introduction.....	3
2. Existing definitions.....	6
3. Test Considerations.....	7
3.1. Paths.....	7
3.1.1. Path.....	7
3.1.2. Working Path.....	8
3.1.3. Primary Path.....	8
3.1.4. Protected Primary Path.....	8
3.1.5. Backup Path.....	9
3.1.6. Standby Backup Path.....	10
3.1.7. Dynamic Backup Path.....	10
3.1.8. Disjoint Paths.....	10
3.1.9. Point of Local repair (PLR).....	11
3.1.10. Shared Risk Link Group (SRLG).....	11
3.2. Protection Mechanisms.....	12
3.2.1. Link Protection.....	12
3.2.2. Node Protection.....	12
3.2.3. Path Protection.....	12
3.2.4. Backup Span.....	13
3.2.5. Local Link Protection.....	13
3.2.6. Redundant Node Protection.....	14
3.2.7. State Control Interface.....	14
3.2.8. Protected Interface.....	15
3.3. Protection Switching.....	15
3.3.1. Protection Switching System.....	15
3.3.2. Failover Event.....	15
3.3.3. Failure Detection.....	16
3.3.4. Failover.....	17
3.3.5. Restoration.....	17
3.3.6. Reversion.....	18
3.4. Nodes.....	18
3.4.1. Protection-Switching Node.....	18
3.4.2. Non-Protection Switching Node.....	19
3.4.3. Headend Node.....	19
3.4.4. Backup Node.....	19
3.4.5. Merge Node.....	20
3.4.6. Primary Node.....	20
3.4.7. Standby Node.....	21
3.5. Benchmarks.....	21
3.5.1. Failover Packet Loss.....	21
3.5.2. Reversion Packet Loss.....	22
3.5.3. Failover Time.....	22
3.5.4. Reversion Time.....	23
3.5.5. Additive Backup Delay.....	23
3.6. Failover Time Calculation Methods.....	24
3.6.1. Time-Based Loss Method.....	24
3.6.2. Packet-Loss Based Method.....	25
3.6.3. Timestamp-Based Method.....	25
4. Acknowledgments.....	26
5. IANA Considerations.....	26
6. Security Considerations.....	26
7. References.....	26
8. Authors' Addresses.....	27

1. Introduction

The IP network layer provides route convergence to protect data traffic against planned and unplanned failures in the internet. Fast convergence times are critical to maintain reliable network connectivity and performance. Convergence Events [6] are recognized at the IP Layer so that Route Convergence [6] occurs. Technologies that function at sub-IP layers can be enabled to provide further protection of IP traffic by providing the failure recovery at the sub-IP layers so that the outage is not observed at the IP-layer. Such sub-IP protection technologies include, but are not limited to, High Availability (HA) stateful failover, Virtual Router Redundancy Protocol (VRRP) [8], Automatic Link Protection (APS) for SONET/SDH, Resilient Packet Ring (RPR) for Ethernet, and Fast Reroute for Multi-Protocol Label Switching (MPLS-FRR) [9].

1.1 Scope

Benchmarking terminology was defined for IP-layer convergence in [6]. Different terminology and methodologies specific to benchmarking sub-IP layer protection mechanisms are required. The metrics for benchmarking the performance of sub-IP protection mechanisms are measured at the IP layer, so that the results are always measured in reference to IP and independent of the specific protection mechanism being used. The purpose of this document is to provide a single terminology for benchmarking sub-IP protection mechanisms.

A common terminology for Sub-IP layer protection mechanism benchmarking enables different implementations of a protection mechanism to be benchmarked and evaluated. In addition, implementations of different protection mechanisms can be benchmarked and evaluated. It is intended that there can exist unique methodology documents for each sub-IP protection mechanism based upon this common terminology document. The terminology can be applied to methodologies that benchmark sub-IP protection mechanism performance with a single stream of traffic or multiple streams of traffic. The traffic flow may be uni-directional or bi-directional as to be indicated in the methodology.

1.2 General Model

The sequence of events to benchmark the performance of Sub-IP Protection Mechanisms is as follows:

1. Failover Event - Primary Path fails
2. Failure Detection- Failover Event is detected
3. Failover - Backup Path becomes the Working Path due to Failover Event
4. Restoration - Primary Path recovers from a Failover Event
5. Reversion (optional) - Primary Path becomes the Working Path

These terms are further defined in this document.

Figures 1 through 5 show models that MAY be used when benchmarking Sub-IP Protection mechanisms, which MUST use a Protection Switching System that consists of a minimum of two Protection-Switching Nodes, an Ingress Node known as the Headend Node and an Egress Node known as the Merge Node. The Protection Switching System MUST include either a Primary Path and Backup Path, as shown in Figures 1 through 4, or a Primary Node and Standby Node, as shown in Figure 5. A Protection Switching System may provide link protection, node protection, path protection, local link protection, and high availability, as shown in Figures 1 through 5 respectively. A Failover Event occurs along the Primary Path or at the Primary Node. The Working Path is the Primary Path prior to the Failover Event and the Backup Path after the Failover Event. A Tester is set outside the two paths or nodes as it sends and receives IP traffic along the Working Path. The tester MUST record the IP packet sequence numbers, departure time, and arrival time so that the metrics of Failover Time, Additive Latency, Packet Reordering, Duplicate Packets, and Reversion Time can be measured. The Tester may be a single device or a test system. If Reversion is supported then the Working Path is the Primary Path after Restoration (Failure Recovery) of the Primary Path.

Link Protection, as shown in Figure 1, provides protection when a Failover Event occurs on the link between two nodes along the Primary Path. Node Protection, as shown in Figure 2, provides protection when a Failover Event occurs at a Node along the Primary Path. Path Protection, as shown in Figure 3, provides protection for link or node failures for multiple hops along the Primary Path. Local Link Protection, as shown in Figure 4, provides Sub-IP Protection of a link between two nodes, without a Backup Node. An example of such a Sub-IP Protection mechanism is SONET APS. High Availability Protection, as shown in Figure 5, provides protection of a Primary Node with a redundant Standby Node. State Control is provided between the Primary and Standby Nodes. Failure of the Primary Node is detected at the Sub-IP layer to force traffic to switch to the Standby Node, which has state maintained for zero or minimal packet loss.

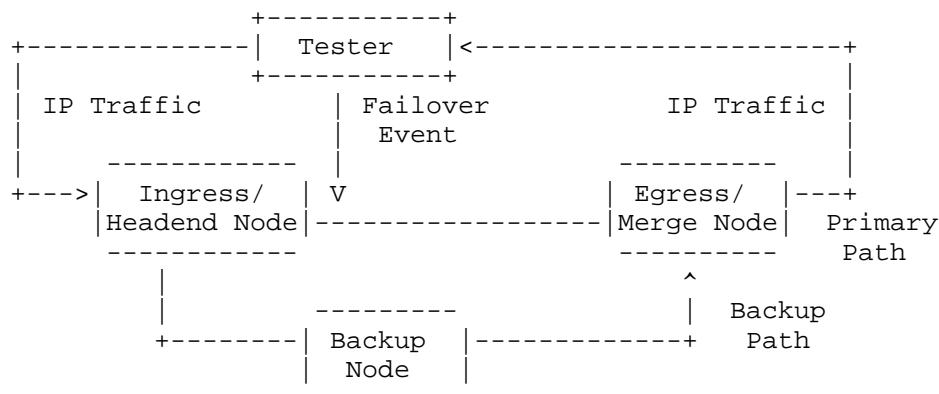


Figure 1. System Under Test (SUT) for Sub-IP Link Protection

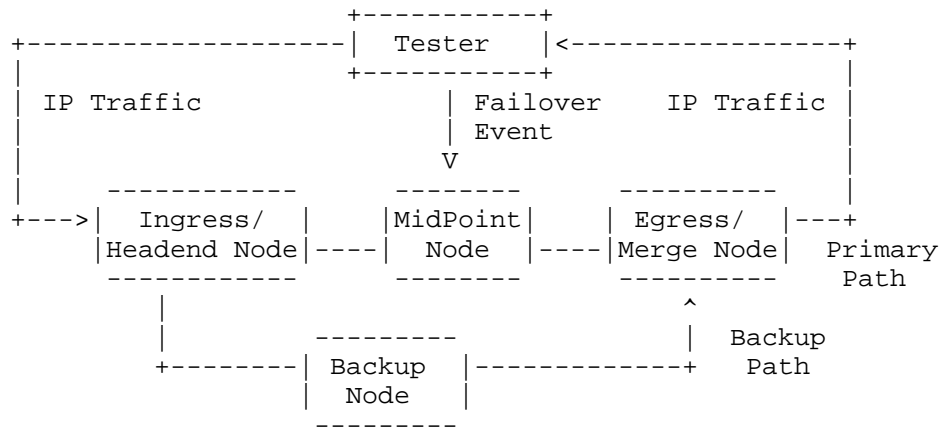


Figure 2. System Under Test (SUT) for Sub-IP Node Protection

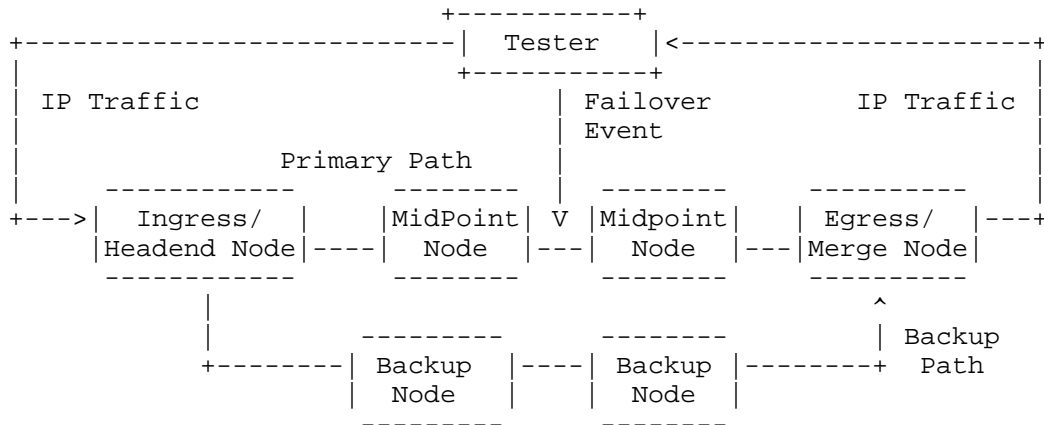


Figure 3. System Under Test (SUT) for Sub-IP Path Protection

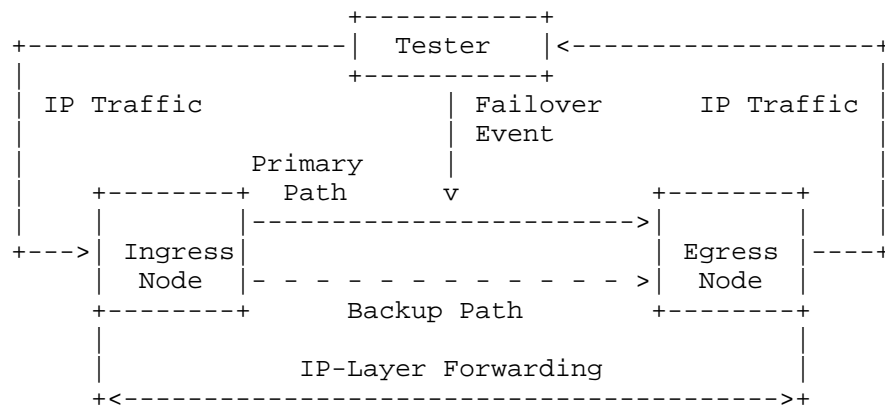


Figure 4. System Under Test (SUT) for Sub-IP Local Link Protection

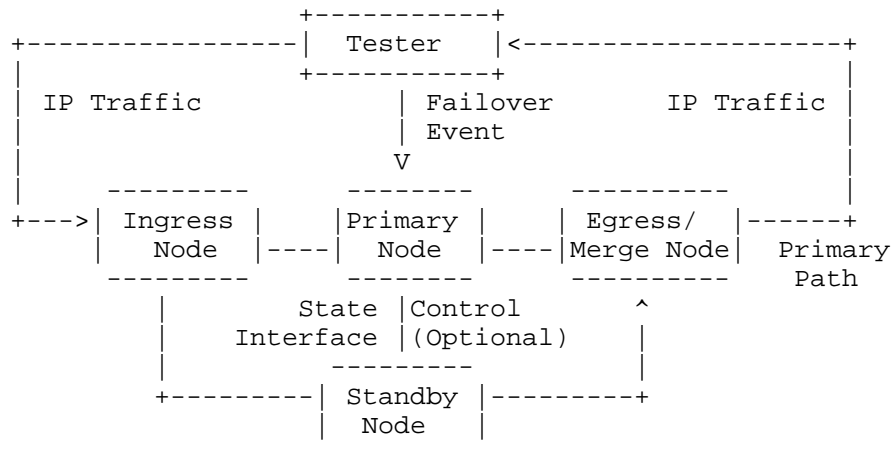


Figure 5. System Under Test (SUT) for Sub-IP Redundant Node Protection

Some protection switching technologies may use a series of steps that differ from the general model. The specific differences SHOULD be highlighted in each technology-specific methodology. Note that some protection switching technologies are endowed with the ability to re-optimize the working path after a node or link failure.

2. Existing definitions

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

Latency	[Ref.[2], section 3.8]
Frame Loss Rate	[Ref.[2], section 3.6]
Throughput	[Ref.[2], section 3.17]
Device Under Test (DUT)	[Ref.[3], section 3.1.1]
System Under Test (SUT)	[Ref.[3], section 3.1.2]
Offered Load	[Ref.[3], section 3.5.2]
Out-of-order Packet	[Ref.[4], section 3.3.2]
Duplicate Packet	[Ref.[4], section 3.3.3]
Forwarding Delay	[Ref.[4], section 3.2.4]
Jitter	[Ref.[4], section 3.2.5]
Packet Loss	[Ref.[6], Section 3.5]
Packet Reordering	[Ref.[7], section 3.3]

This document has the following frequently used acronyms:

DUT	Device Under Test
SUT	System Under Test

This document adopts the definition format in Section 2 of RFC 1242 [2]. Terms defined in this document are capitalized when used within this document.

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 [5].

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. Test Considerations

3.1. Paths

3.1.1 Path

Definition:

A unidirectional sequence of nodes, $\langle R1, \dots, Rn \rangle$, and links $\langle L12, \dots, L(n-1)n \rangle$ with the following properties:

- a. $R1$ is the ingress node and forwards IP packets, which input into DUT/SUT, to $R2$ as sub-IP frames over link $L12$.
- b. Ri is a node which forwards data frames to $R(i+1)$ over Link $Li(i+1)$ for all i , $1 < i < n-1$, based on information in the sub-IP layer.
- c. Rn is the egress node and it outputs sub-IP frames from DUT/SUT as IP packets. $L(n-1)n$ is the link between the $R(n-1)$ and Rn .

Discussion:

The path is defined in the sub-IP layer in this document, unlike an IP path in RFC 2026 [1]. One path may be regarded as being equivalent to one IP link between two IP nodes, i.e., $R1$ and Rn . The two IP nodes may have multiple paths for protection. A packet will travel on only one path between the nodes. Packets belonging to a microflow [10] will traverse one or more paths. The path is unidirectional. For example, the link between $R1$ and $R2$ in the direction from $R1$ to $R2$ is $L12$. For traffic flowing in the reverse direction from $R2$ to $R1$, the link is $L21$. Example paths are the SONET/SDH path and the label switched path for MPLS.

Measurement units:

n/a

Issues:

"A bidirectional path", which transmits traffic in both directions along the same nodes, consists of two unidirectional paths. Therefore, the two unidirectional paths belonging to "one bidirectional path" will be treated independently when benchmarking for "a bidirectional path".

See Also:

Working Path
Primary Path
Backup Path

3.1.2. Working Path

Definition:

The path that the DUT/SUT is currently using to forward packets.

Discussion:

A Primary Path is the Working Path before occurrence of a Failover Event. A Backup Path shall become the Working Path after a Failover Event.

Measurement units:

n/a

Issues:

See Also:

Path
Primary Path
Backup Path

3.1.3. Primary Path

Definition:

The preferred point to point path for forwarding traffic between two or more nodes.

Discussion:

The Primary Path is the Path that traffic traverses prior to a Failover Event.

Measurement units:

n/a

Issues:

None

See Also:

Path
Failover Event

3.1.4. Protected Primary Path

Definition:

A Primary Path that is protected with a Backup Path.

Discussion:

A Protected Primary Path must include at least one Protection Switching Node.

Measurement units:
n/a

Issues: None

See Also:
Path
Primary Path

3.1.5. Backup Path

Definition:

A path that exists to carry data traffic only if a Failover Event occurs on a Primary Path.

Discussion:

The Backup Path shall become the Working Path upon a Failover Event. A Path may have one or more Backup Paths. A Backup Path may protect one or more Primary Paths. There are various types of Backup Paths:

- a. dedicated recovery Backup Path (1+1) or (1:1), which has 100% redundancy for a specific ordinary path,
- b. shared Backup Path (1:N), which is dedicated to the protection for more than one specific Primary Path
- c. associated shared Backup Path (M:N) for which a specific set of Backup Paths protects a specific set of more than one Primary Path.

A Backup Path may be signaled or un signaled. The Backup Path must be created prior to the Failover Event. The backup path generally originates at the point of local repair (PLR), and terminates at a node along a primary path.

Measurement units:
n/a

Issues:

See Also:
Path
Working Path
Primary Path

3.1.6. Standby Backup Path

Definition:

A Backup Path that is established prior to a Failover Event to protect a Primary Path.

Discussion:

The Standby Backup Path and Dynamic Backup Path provide protection, but are established at different times.

Measurement units: n/a

Issues: None

See Also:

- Backup Path
- Primary Path
- Failover Event

3.1.7. Dynamic Backup Path

Definition:

A Backup Path that is established upon occurrence of a Failover Event.

Discussion:

The Standby Backup Path and Dynamic Backup Path provide protection, but are established at different times.

Measurement units: n/a

Issues: None

See Also:

- Backup Path
- Standby Backup Path
- Failover Event

3.1.8. Disjoint Paths

Definition:

A pair of paths that do not share a common link or nodes.

Discussion:

Two paths are disjoint if they do not share a common node or link other than the ingress and egress.

Measurement units: n/a

Issues: None

See Also:

- Path
- Primary Path
- SRLG

3.1.9. Point of Local Repair (PLR)

Definition:

A node capable of Failover along the Primary Path that is also the ingress node for the Backup Path to protect another node or link.

Discussion:

Any node along the Primary Path from the ingress node to the penultimate node may be a PLR. The PLR may use a single Backup Path for protecting one or more Primary Paths. There can be multiple PLRs along a Primary Path. The PLR must be an ingress to a Backup Path. The PLR can be any node along the Primary Path except the egress node of the Primary Path. The PLR may simultaneously be a Headend Node when it is serving the role as ingress to the Primary Path and the Backup Path. If the PLR is also the Headend Node, then the Backup Path is a Disjoint Path from the ingress to the Merge Node.

Measurement units: n/a

Issues: None

See Also:

- Primary Path
- Backup Path
- Failover

3.1.10. Shared Risk Link Group (SRLG)

Definition:

SRLG is a set of links which share the same risk (physical or logical) within a network.

Discussion:

SRLG is considered the set of links to be avoided when the primary and secondary paths are considered disjoint. The SRLG will fail as a group if the shared resource (physical or anything abstract such as software version) fails.

Measurement units: n/a

Issues: None

See Also:

- Path
- Primary Path

3.2. Protection

3.2.1. Link Protection

Definition:

A Backup Path that is signaled to at least one Backup Node to protect for failure of interfaces and links along a Primary Path.

Discussion:

Link Protection may or may not protect the entire Primary Path. Link protection is shown in Figure 1.

Measurement units: n/a

Issues: None

See Also:

Primary Path
Backup Path

3.2.2. Node Protection

Definition:

A Backup Path that is signaled to at least one Backup Node to protect for failure of interfaces, links, and nodes along a Primary Path.

Discussion:

Node Protection may or may not protect the entire Primary Path. Node Protection also provides Link Protection. Node Protection is shown in Figure 2.

Measurement units: n/a

Issues: None

See Also:

Link Protection

3.2.3. Path Protection

Definition:

A Backup Path that is signaled to at least one Backup Node to provide protection along the entire Primary Path.

Discussion:

Path Protection provides Node Protection and Link Protection for every node and link along the Primary Path. A Backup Path providing Path Protection may have the same ingress node as the Primary Path. Path Protection is shown in Figure 3.

Measurement units: n/a

Issues: None

See Also:

- Primary Path
- Backup Path
- Node Protection
- Link protection

3.2.4. Backup Span

Definition:

The number of hops used by a Backup Path.

Discussion:

The Backup Span is an integer obtained by counting the number of nodes along the Backup Path.

Measurement units:

number of nodes

Issues:

None

See Also:

- Primary Path
- Backup Path

3.2.5. Local Link Protection

Definition:

A Backup Path that is a redundant path between two nodes which does not use a Backup Node.

Discussion:

Local Link Protection must be provided as a Backup Path between two nodes along the Primary Path without the use of a Backup Node. Local Link Protection is provided by Protection Switching Systems such as SONET APS. Local Link Protection is shown in Figure 4.

Measurement units: None

Issues: None

See Also:

- Backup Path
- Backup Node

3.2.6. Redundant Node Protection

Definition:

A Protection Switching System with a Primary Node protected by a Standby Node along the Primary Path.

Discussion:

Redundant Node Protection is provided by Protection Switching Systems such as VRRP and HA. The protection mechanisms occur at Sub-IP layers to switch traffic from a Primary Node to Backup Node upon a Failover Event at the Primary Node. Traffic continues to traverse the Primary Path through the Standby Node. The failover may be stateful, in which the state information may be exchanged in-band or over an out-of-band state control interface. The Standby Node may be active or passive. Redundant Node Protection is shown in Figure 5.

Measurement units: None

Issues: None

See Also:

Primary Path
Primary Node
Standby Node

3.2.7. State Control Interface

Definition:

An out-of-band control interface used to exchange state information between the Primary Node and Standby Node.

Discussion:

The State Control Interface may be used for Redundant Node Protection. The State Control Interface should be out-of-band. It is possible to have Redundant Node Protection in which there is no state control or state control is provided in-band. The State Control Interface between the Primary and Standby Node may be one or more hops.

Measurement units: None

Issues: None

See Also:

Primary Node
Standby Node

3.2.8. Protected Interface

Definition:

An interface along the Primary Path that is protected by a Backup Path.

Discussion:

A Protected Interface is an interface protected by a Protection Switching System that provides Link Protection, Node Protection, Path Protection, Local Link Protection, and Redundant Node Protection.

Measurement units: None

Issues: None

See Also:

Primary Path
Backup Path

3.3. Protection Switching

3.3.1. Protection Switching System

Definition:

A DUT/SUT that is capable of Failure Detection and Failover from a Primary Path to a Backup Path or Standby Node when a Failover Event occurs.

Discussion:

The Protection Switching System must include either a Primary Path and Backup Path, as shown in Figures 1 through 4, or a Primary Node and Standby Node, as shown in Figure 5. The Backup Path may be a Standby Backup Path or a dynamic Backup Path. The Protection Switching System includes the mechanisms for both Failure Detection and Failover.

Measurement units: n/a

Issues: None

See Also:

Primary Path
Backup Path
Failover

3.3.2. Failover Event

Definition:

The occurrence of a planned or unplanned action in the network that results in a change in the Path that data traffic traverses.

Discussion:

Failover Events include, but are not limited to, link failure and router failure. Routing changes are considered Convergence Events [6] and are not Failover Events. This restricts Failover Events to sub-IP layers. Failover may be at the PLR or at the ingress. If the failover is at the ingress it is generally on a disjoint path from the ingress to egress.

Failover Events may results from failures such as link failure or router failure. The change in path after Failover may have a Backup Span of one or more nodes. Failover Events are distinguished from routing changes and Convergence Events [6] by the detection of the failure and subsequent protection switching at a sub-IP layer. Failover occurs at a Point of Local Repair (PLR) or Primary Node.

Measurement units:

n/a

Issues: None

See Also:

Path
Failure Detection
Disjoint Path

3.3.3. Failure Detection

Definition:

The process to identify at a sub-IP layer a Failover Event at a Primary Node or along the Primary Path.

Discussion:

Failure Detection occurs at the Primary Node or ingress node of the Primary Path. Failure Detection occurs via a sub-IP mechanism such as detection of a link down event or timeout for receipt of a control packet. A failure may be completely isolated. A failure may affect a set of links which share a single SRLG (e.g. port with many sub-interfaces). A failure may affect multiple links that are not part of SRLG.

Measurement units: n/a

Issues:

See Also:

Primary Path

3.3.4. Failover

Definition:

The process to switch data traffic from the protected Primary Path to the Backup Path upon Failure Detection of a Failover Event.

Discussion:

Failover to a Backup Path provides Link Protection, Node Protection, or Path Protection. Failover is complete when Packet Loss [6], Out-of-order Packets [4], and Duplicate Packets [4] are no longer observed. Forwarding Delay [4] may continue to be observed.

Measurement units:

n/a

Issues:

See Also:

- Primary Path
- Backup Path
- Failover Event

3.3.5. Restoration

Definition:

The state of failover recovery in which the Primary Path has recovered from a Failover Event, but is not yet forwarding packets because the Backup Path remains the Working Path.

Discussion:

Restoration must occur while the Backup Path is the Working Path. The Backup Path is maintained as the Working Path during Restoration. Restoration produces a Primary Path that is recovered from failure, but is not yet forwarding traffic. Traffic is still being forwarded by the Backup Path functioning as the Working Path.

Measurement units:

n/a

Issues:

See Also:

- Primary Path
- Failover Event
- Failure Recovery
- Working Path
- Backup Path

3.3.6. Reversion

Definition:

The state of failover recovery in which the Primary Path has become the Working Path so that it is forwarding packets.

Discussion:

Protection Switching Systems may or may not support Reversion. Reversion, if supported, must occur after Restoration. Packet forwarding on the Primary Path resulting from Reversion may occur either fully or partially over the Primary Path. A potential problem with Reversion is the discontinuity in end to end delay when the Forwarding Delays [4] along the Primary Path and Backup Path are different, possibly causing Out of Order Packets [4], Duplicate Packets [4], and increased Jitter [4].

Measurement units: n/a

Issues: None

See Also:

Protection Switching System
Working Path
Primary Path

3.4. Nodes

3.4.1. Protection-Switching Node

Definition:

A node that is capable of participating in a Protection Switching System.

Discussion:

The Protection Switching Node may be an ingress or egress for a Primary Path or Backup Path, such as used for MPLS Fast Reroute configurations. The Protection Switching Node may provide Redundant Node Protection as a Primary Node in a Redundant chassis configuration with a Standby Node, such as used for VRRP and HA configurations.

Measurement units:

n/a

Issues:

See Also:

Protection Switching System

3.4.2. Non-Protection Switching Node

Definition:

A node that is not capable of participating in a Protection Switching System, but may exist along the Primary Path or Backup Path.

Discussion:

Measurement units:

n/a

Issues:

See Also:

Protection Switching System
Primary Path
Backup Path

3.4.3. Headend Node

Definition:

The ingress node of the Primary Path.

Discussion:

The Headend Node may also be a PLR when it is serving in the dual role as the ingress to the Backup Path.

Measurement units: n/a

Issues:

See Also:

Primary Path
Point of Local Repair (PLR)
Failover

3.4.4. Backup Node

Definition:

A node along the Backup Path.

Discussion:

The Backup Node can be any node along the Backup Path. There may be one or more Backup Nodes along the Backup Path. A Backup Node may be the ingress, mid-point, or egress of the Backup Path. If the Backup Path has only one Backup Node, then that Backup Node is the ingress and egress of the Backup Path.

Measurement units: n/a

Issues:

See Also:

Backup Path

3.4.5. Merge Node

Definition:

A node along the Primary Path where Backup Path terminates.

Discussion:

The Merge Node can be any node along the Primary Path except the ingress node of the Primary Path. There can be multiple Merge Nodes along a Primary Path. A Merge Node can be the egress node for a single or multiple Backup Paths. The Merge Node must be the egress to the Backup Path. The Merge Node may also be the egress of the Primary Path or Point of Local Repair (PLR).

Measurement units:

n/a

Issues:

See Also:

Primary Path

Backup Path

PLR

Failover

3.4.6. Primary Node

Definition:

A node along the Primary Path that is capable of Failover to a redundant Standby Node.

Discussion:

The Primary Node may be used for Protection Switching Systems that provide Redundant Node Protection, such as VRRP and HA

Measurement units: n/a

Issues:

See Also:

Protection Switching System

Redundant Node Protection

Standby Node

3.4.7. Standby Node

Definition:

A redundant node to a Primary Node that forwards traffic along the Primary Path upon Failure Detection of the Primary Node.

Discussion:

The Standby Node must be used for Protection Switching Systems that provide Redundant Node Protection, such as VRRP and HA. The Standby Node must provide protection along the same Primary Path. If the failover is to a Disjoint Path then it is a Backup Node. The Standby Node may be configured for 1:1 or N:1 protection.

The communication between the Primary Node and Standby Node may be in-band or across an out-of-band State Control interface. The Standby Node may be geographically dispersed from the Primary Node. When geographically dispersed, the number of hops of separation may increase failover time.

The Standby Node may be passive or active. The Passive Standby Node is not offered traffic and does not forward traffic until Failure Detection of the Primary Node. Upon Failure Detection of the Primary Node, traffic offered to the Primary Node is instead offered to the Passive Standby Node. The Active Standby Node is offered traffic and forwards traffic along the Primary Path while the Primary Node is also active. Upon Failure Detection of the Primary Node, traffic offered to the Primary Node is switched to the Active Standby Node.

Measurement units: n/a

Issues:

See Also:

Primary Node
State Control Interface

3.5. Benchmarks

3.5.1. Failover Packet Loss

Definition:

The amount of packet loss produced by a Failover Event until Failover completes, where the measurement begins when the last unimpaired packet is received by the Tester on the Protected Primary Path and ends when the first unimpaired packet is received by the Tester on the Backup Path.

Discussion:

Packet loss can be observed as a reduction of forwarded traffic from the maximum forwarding rate. Failover Packet Loss includes packets that were lost, reordered, or delayed. Failover Packet Loss may reach 100% of the offered load.

Measurement units:

Number of Packets

Issues: None

See Also:

Failover Event
Failover

3.5.2. Reversion Packet Loss

Definition:

The amount of packet loss produced by Reversion, where the measurement begins when the last unimpaired packet is received by the Tester on the Backup Path and ends when the first unimpaired packet is received by the Tester on the Protected Primary Path .

Discussion:

Packet loss can be observed as a reduction of forwarded traffic from the maximum forwarding rate. Reversion Packet Loss includes packets that were lost, reordered, or delayed. Reversion Packet Loss may reach 100% of the offered load.

Measurement units: Number of Packets

Issues: None

See Also:

Reversion

3.5.3. Failover Time

Definition:

The amount of time it takes for Failover to successfully complete.

Discussion:

Failover Time can be calculated using the Time-Based Loss Method (TBLM), Packet-Loss Based Method (PLBM), or Timestamp-Based Method (TBM). It is RECOMMENDED that the TBM is used.

Measurement units:
 milliseconds

Issues: None

See Also:
 Failover
 Failover Time
 Time-Based Loss Method (TBLM)
 Packet-Loss Based Method (PLBM)
 Timestamp-Based Method (TBM)

3.5.4. Reversion Time

Definition:
The amount of time it takes for Reversion to complete so that the Primary Path is restored as the Working Path.

Discussion:
Reversion Time can be calculated using the Time-Based Loss Method (TBLM), Packet-Loss Based Method (PLBM), or Timestamp-Based Method (TBM). It is RECOMMENDED that the TBM is used.

Measurement units:
 milliseconds

Issues: None

See Also:
 Reversion
 Primary Path
 Working Path
 Reversion Packet Loss
 Time-Based Loss Method (TBLM)
 Packet-Loss Based Method (PLBM)
 Timestamp-Based Method (TBM)

3.5.5. Additive Backup Delay

Definition:
The amount of increased Forwarding Delay [4] resulting from data traffic traversing the Backup Path instead of the Primary Path.

Discussion:
Additive Backup Delay is calculated using Equation 1 as shown below:

(Equation 1)
Additive Backup Delay =
 Forwarding Delay(Backup Path) -
 Forwarding Delay(Primary Path).

Measurement units:
 milliseconds

Issues:
Additive Backup Latency may be a negative result.
This is theoretically possible, but could be indicative
of a sub-optimum network configuration .

See Also:
 Primary Path
 Backup Path
 Primary Path Latency
 Backup Path Latency

3.6 Failover Time Calculation Methods

The following Methods may be assessed on a per-flow basis using at least 16 flows spread over the routing table (more flows is better). Otherwise, the impact of a prefix-dependency in the implementation of a particular protection technology could be missed. However, the test designer must be aware of the number of packets per second sent to each prefix, as this establishes sampling of the path and the time resolution for measurement of Failover time on a per-flow basis.

3.6.1 Time-Based Loss Method (TBLM)

Definition:
The method to calculate Failover Time (or Reversion Time) using a time scale on the Tester to measure the interval of Failover Packet Loss.

Discussion:
The Tester must provide statistics which show the duration of failure on a time scale based on occurrence of packet loss on a time scale. This is indicated by the duration of non-zero packet loss. The TBLM includes failure detection time and time for data traffic to begin traversing the Backup Path. Failover Time and Reversion Time are calculated using the TBLM as shown in Equation 2:

(Equation 2)

(Equation 2a)

$$\text{TBLM Failover Time} = \text{Time(Failover)} - \text{Time(Failover Event)}$$

(Equation 2b)

$$\text{TBLM Reversion Time} = \text{Time(Reversion)} - \text{Time(Restoration)}$$

Where as Time(Failover) = Time on the tester at the receipt of the first unimpaired packet at egress node after the backup path became the working path

$\text{Time(Failover Event)}$ = Time on the tester at the receipt of the last unimpaired packet at egress node on the primary path before failure

Measurement units:
 milliseconds

Issues:
 None

See Also:

Failover
Packet-Loss Based Method

3.6.2 Packet-Loss Based Method (PLBM)

Definition:

The method used to calculate Failover Time (or Reversion Time) from the amount of Failover Packet Loss.

Discussion:

PLBM includes failure detection time and time for data traffic to begin traversing the Backup Path. Failover Time can be calculated using PLBM from the amount Failover Packet Loss as shown below in Equation 3. Note: If traffic is sent to more than 1 destination, PLBM gives the average loss over the measured destinations

(Equation 3)

(Equation 3a)

$$\text{PLBM Failover Time} = \frac{(\text{Number of packets lost})}{(\text{Offered Load rate})} * 1000$$

(Equation 3b)

$$\text{PLBM Restoration Time} = \frac{(\text{Number of packets lost})}{(\text{Offered Load rate})} * 1000$$

Units are packets/(packets/second) = seconds

Measurement units:

milliseconds

Issues:

None

See Also:

Failover
Time-Based Loss Method

3.6.3 Timestamp-Based Method (TBM)

Definition:

The method to calculate Failover Time (or Reversion Time) using a time scale to quantify the interval between unimpaired packets arriving in the test stream.

Discussion:

The purpose of this method is to quantify the duration of failure or reversion on a time scale based on the observation of unimpaired packets. The TBM is calculated from Equation 2 with the values obtained from the timestamp in the packet payload, rather than from the Tester clock as is used for the values when using the TBLM.

Unimpaired packets are normal packets that are not lost, reordered, or duplicated. A reordered packet is defined in

[10, section 3.3]. A duplicate packet is defined in [4, section 3.3.3]. A lost packet is defined in [7, Section 3.5]. Unimpaired packets may be detected by checking a sequence number in the payload, where the sequence number equals the next expected number for an unimpaired packet. A sequence gap or sequence reversal indicates impaired packets.

For calculating Failover Time, the TBM includes failure detection time and time for data traffic to begin traversing the Backup Path. For calculating Reversion Time, the TBM includes Reversion Time and time for data traffic to begin traversing the Primary Path.

Measurement units:
 milliseconds

Issues: None

See Also:
 Failover
 Failover Time
 Reversion
 Reversion Time

4. Acknowledgements

We would like thank the BMWG and particularly Al Morton and Curtis Villamizar for their reviews, comments, and contributions to this work.

5. IANA Considerations

This document requires no IANA considerations.

6. 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.

7. References

7.1. Normative References

- [1] Bradner, S., "The Internet Standards Process -- Revision 3", RFC 2026, October 1996.
- [2] Bradner, S., Editor, "Benchmarking Terminology for Network Interconnection Devices", RFC 1242, July 1991.
- [3] Mandeville, R., "Benchmarking Terminology for LAN Switching Devices", RFC 2285, February 1998.
- [4] Poretsky, S., et al., "Terminology for Benchmarking Network-layer Traffic Control Mechanisms", RFC 4689, November 2006.
- [5] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", RFC 2119, July 1997.
- [6] Poretsky, S., Imhoff, B., "Benchmarking Terminology for IGP Convergence", draft-ietf-bmwg-igp-dataplane-conv-term-21, work in progress, May 2010.
- [7] Morton, A., et al, "Packet Reordering Metrics", RFC 4737, November 2006.
- [8] Hinden, R., "Virtual Router Redundancy Protocol", RFC 5798, March 2010.

7.2. Informative References

- [9] Pan., P. et al, "Fast Reroute Extensions to RSVP-TE for LSP Paths", RFC 4090, May 2005.
- [10] Nichols, K., et al, "Definition of the Differentiated Services Field (DS Field) in the IPv4 and IPv6 Headers", RFC 2474, December 1998.

8. 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

Rajiv Papneja
Isocore
12359 Sunrise Valley Drive
Reston, VA 22102
USA
Phone: +1 703 860 9273
Email: rpapneja@isocore.com

Jay Karthik
Cisco Systems
300 Beaver Brook Road
Boxborough, MA 01719
USA
Phone: +1 978 936 0533
Email: jkarthik@cisco.com

Samir Vapiwala
Cisco System
300 Beaver Brook Road
Boxborough, MA 01719
USA
Phone: +1 978 936 1484
Email: svapiwal@cisco.com

Benchmarking Methodology Working Group
Internet-Draft
Intended status: Informational
Expires: May 16, 2015

C. Davids
Illinois Institute of Technology
V. Gurbani
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.

Status of this Memo

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

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

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

This Internet-Draft will expire on May 16, 2015.

Copyright Notice

Copyright (c) 2014 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Terminology	4
2. Introduction	4
3. Benchmarking Topologies	5
4. Test Setup Parameters	7
4.1. Selection of SIP Transport Protocol	7
4.2. Connection-oriented Transport Management	7
4.3. Signaling Server	8
4.4. Associated Media	8
4.5. Selection of Associated Media Protocol	8
4.6. Number of Associated Media Streams per SIP Session	8
4.7. Codec Type	8
4.8. Session Duration	8
4.9. Attempted Sessions per Second (sps)	9
4.10. Benchmarking algorithm	9
5. Reporting Format	11
5.1. Test Setup Report	11
5.2. Device Benchmarks for session setup	12
5.3. Device Benchmarks for registrations	12
6. Test Cases	13
6.1. Baseline Session Establishment Rate of the test bed	13
6.2. Session Establishment Rate without media	13
6.3. Session Establishment Rate with Media not on DUT	13
6.4. Session Establishment Rate with Media on DUT	14
6.5. Session Establishment Rate with TLS Encrypted SIP	14
6.6. Session Establishment Rate with IPsec Encrypted SIP	15
6.7. Registration Rate	15
6.8. Re-Registration Rate	16
7. IANA Considerations	16
8. Security Considerations	16
9. Acknowledgments	17
10. References	17
10.1. Normative References	17
10.2. Informative References	17
Appendix A. R Code Component to simulate benchmarking algorithm	18
Authors' Addresses	20

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.

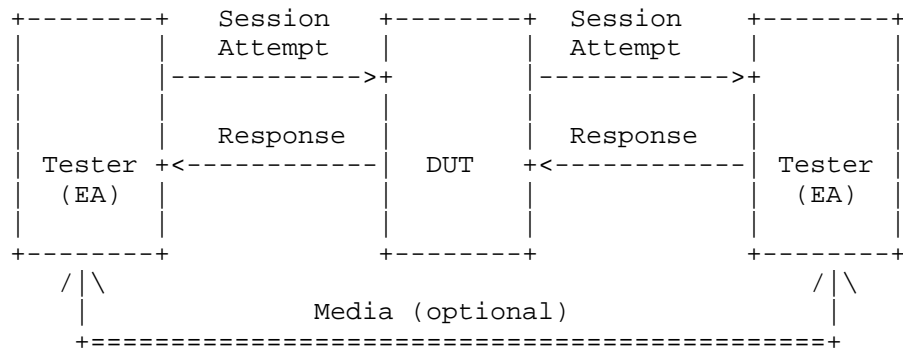


Figure 1: DUT as an intermediary, end-to-end media

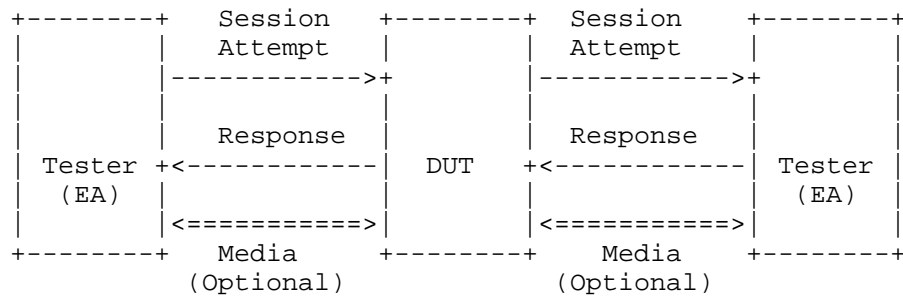


Figure 2: DUT as an intermediary forwarding media

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

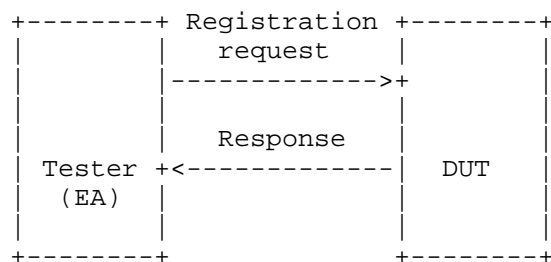


Figure 3: Registration and Re-registration tests

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
            }
        }
    }
end proc

```

```

        }
    }
    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 Cerveney 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
Internet-Draft
Intended status: Informational
Expires: May 16, 2015

C. Davids
Illinois Institute of Technology
V. Gurbani
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 is submitted in full conformance with the provisions of BCP 78 and BCP 79.

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

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

This Internet-Draft will expire on May 16, 2015.

Copyright Notice

Copyright (c) 2014 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Terminology	4
2. Introduction	4
2.1. Scope	6
3. Term Definitions	7
3.1. Protocol Components	7
3.1.1. Session	7
3.1.2. Signaling Plane	8
3.1.3. Media Plane	8
3.1.4. Associated Media	9
3.1.5. Overload	9
3.1.6. Session Attempt	10
3.1.7. Established Session	10
3.1.8. Session Attempt Failure	11
3.2. Test Components	11
3.2.1. Emulated Agent	11
3.2.2. Signaling Server	12
3.2.3. SIP Transport Protocol	12
3.3. Test Setup Parameters	13
3.3.1. Session Attempt Rate	13
3.3.2. Establishment Threshold Time	13
3.3.3. Session Duration	14
3.3.4. Media Packet Size	14
3.3.5. Codec Type	15
3.4. Benchmarks	15
3.4.1. Session Establishment Rate	16
3.4.2. Registration Rate	16
3.4.3. Registration Attempt Rate	17
4. IANA Considerations	17
5. Security Considerations	17
6. Acknowledgments	18
7. References	18
7.1. Normative References	18
7.2. Informational References	19
Authors' Addresses	19

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].

- o 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.
- o 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.
- o 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 Cervený 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
Internet-Draft
Intended status: Informational
Expires: November 24, 2010

V. Manral, Ed.
IPInfusion Inc.
May 23, 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 beginning 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].

Status of this Memo

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

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

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

This Internet-Draft will expire on November 24, 2010.

Copyright Notice

Copyright (c) 2010 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction	3
2. Challenges in defining benchmarks	3
3. Factors for power consumption	4
3.1. Network Factors affecting power consumption	5
3.2. Device Factors affecting power consumption	5
3.3. Traffic Factors affecting power consumption	6
4. Network Energy Consumption Rate (NECR)	6
5. Network Energy Proportionality Index (NEPI)	6
6. Benchmark details	7
7. IANA Considerations	7
8. Security Considerations	7
9. Acknowledgements	7
10. References	8
10.1. Normative References	8
10.2. Informative References	8
Author's Address	8

1. Introduction

Energy Efficiency is becoming increasingly important in the operation of network infrastructure. Data traffic is exploding at an accelerated rate. Networks provide communication channels that facilitates components of the infrastructures to exchange critical information and are always on. On the other hand, a lot of devices run at very low average utilization 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 be 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 dissipated 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 transceivers/ 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 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.

$$EPIx = (Mx - Ix) / Mx * 100$$

$$EPI = EPI1 + EPI2 + \dots + EPI_n / n$$

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:

