



MPLS Forwarding Application Level Benchmark Specification Implementation Agreement

February 24, 2003
Revision 1.0

Editor(s):

Ganesh Balakrishnan, IBM Corporation, ganeshb@us.ibm.com

Ravi Gunturi, Intel Corporation, ravi.gunturi@intel.com

Copyright © 2002 The Network Processing Forum (NPF). All Rights Reserved.

This document and translations of it may be copied and furnished to others, and derivative works that comment on or otherwise explain it or assist in its implementation may be prepared, copied, published and distributed, in whole or in part, without restriction other than the following, (1) the above copyright notice and this paragraph must be included on all such copies and derivative works, and (2) this document itself may not be modified in any way, such as by removing the copyright notice or references to the NPF, except as needed for the purpose of developing NPF Implementation Agreements.

By downloading, copying, or using this document in any manner, the user consents to the terms and conditions of this notice. Unless the terms and conditions of this notice are breached by the user, the limited permissions granted above are perpetual and will not be revoked by the NPF or its successors or assigns.

THIS DOCUMENT AND THE INFORMATION CONTAINED HEREIN IS PROVIDED ON AN "AS IS" BASIS WITHOUT ANY WARRANTY OF ANY KIND. THE INFORMATION, CONCLUSIONS AND OPINIONS CONTAINED IN THE DOCUMENT ARE THOSE OF THE AUTHORS, AND NOT THOSE OF NPF. THE NPF DOES NOT WARRANT THE INFORMATION IN THIS DOCUMENT IS ACCURATE OR CORRECT. THE NPF DISCLAIMS ALL WARRANTIES, WHETHER EXPRESS, IMPLIED OR STATUTORY, INCLUDING BUT NOT LIMITED THE IMPLIED LIMITED WARRANTIES OF MERCHANTABILITY, TITLE OR FITNESS FOR A PARTICULAR PURPOSE AND NON-INFRINGEMENT OF THIRD PARTY RIGHTS.

The words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in the remainder of this document are to be interpreted as described in the NPF Software API Conventions Implementation Agreement revision 1.0.

For additional information contact:

Network Processing Forum Benchmark Working Group

The Network Processing Forum, 39355 California Street,
Suite 307, Fremont, CA 94538
+1 510 608-5990 phone ♦ info@npforum.org

Table of Contents

1	Revision History	4
2	Scope and Purpose	5
3	Normative References.....	6
4	Acronyms, Abbreviations, and Terminology.....	7
	4.1 Benchmark Terminology	7
	4.2 MPLS Terminology	7
5	MPLS Overview	10
6	Test Configuration	12
	6.1 Reference Design.....	12
	6.2 Test Setup.....	14
7	Benchmark Tests.....	15
	7.1 Data Plane Tests.....	15
Appendix A	Informative Annexes.....	24
	A.1. Traffic Generation.....	25
	A.2. Implementation Kit.....	26

1 Revision History

Revision	Date	Reason for changes
0.1	9/14/01	Initial draft
0.2	4/2/02	Control Plane Tests
0.3	5/24/02	Incorporated comments from 5/17/02 conference call
0.4	6/7/02	Incorporated comments from 6/7/02 conference call and restructured the spec after discussion by authors. Accepted some of the changes made so far. Added text on Frame Sizes, and the number of LSPs per device. Added section on Closed Issues. Changed some of the document wording, and formatting.
0.5	8/4/02	Incorporated comments from July 2002 NPF Benchmark WG meetings.
0.6	9/9/02	Incorporated comments from 9/6/02 NPF Benchmark NP/CP TG Meeting. Fixed some of the cross-references that were broken in the previous revision.
0.7	9/18/02	Changed document title
0.8	10/28/02	Frame size clarification and reporting format changes
0.9	12/03/02	Incorporated Straw Ballot comments
1.0	2/24/03	Added the Implementation Annex as an Appendix, updated references

2 Scope and Purpose

This document defines the methodology used to obtain network processor MPLS application level benchmarks. Specifically, this document describes the tests that may be used to obtain MPLS performance metrics in Ingress, Egress and Transit configurations, the test configurations used and, the formats for reporting the results of the tests. The document is based on the IETF RFC 2544[3], RFC 1242[2], RFC 3031[6] and the NPF IPv4 benchmark (Draft 0.3)[4].

Measurements of the tests described in this document can be used by customers to evaluate the performance of network processors versus their requirements and to compare the performance of different Network Processors.

3 Normative References

The following documents contain provisions, which through reference in this text constitute provisions of this specification. At the time of publication, the editions indicated were valid. All referenced documents are subject to revision, and parties to agreements based on this specification are encouraged to investigate the possibility of applying the most recent editions of the standards indicated below.

1. *Network Processors Benchmarking Framework Document*, Network Processing Forum.
2. *Benchmarking Terminology for Network Interconnect Devices*, IETF RFC 1242.
3. *Benchmarking Methodology for Network Interconnect Devices*, IETF RFC 2544.
4. *Methodology for IPv4 Forwarding Level Benchmarks*, NPF Draft 0.2
5. *Internet Routing Table Statistics*, http://www.merit.edu/ipma/routing_table/
6. *Multiprotocol Label Switching Architecture*, IETF RFC 3031.
7. *Light Reading Tests*,
http://www.lightreading.com/document.asp?doc_id=4009&page_number=11
8. *Internet IP Packet Size Distribution*, <http://www.nlanr.net/NA/Learn/Gm/pktsizes.html>
9. Davie, B., and Rekhter, Y. *MPLS: Technology and Applications*. San Francisco: Morgan Kaufmann, 2000.
10. *MPLS Label Stack Encoding*, IETF RFC 3032
11. NPF MPLS Forwarding Application Level Benchmark Implementation Kit
12. Mae West Routing Table Snapshot (mae_west.txt contained in [11])
13. Script for Generating Benchmark Routing Table Subsets (parseMaeWest.tcl contained in [11])
14. NPF Reporting Template for the MPLS Forwarding Application Level Benchmark

4 Acronyms, Abbreviations, and Terminology

This section defines the terms used in the document. The goal is to define a specific terminology that every network processor vendor can use to measure and report the results of the benchmark tests described later on in this document. The benchmarking terminology is based on the IETF benchmarking terminology for network interconnection devices described in RFC 1242[2]. The MPLS terminology is based on the IETF RFC 3031[6].

4.1 Benchmark Terminology

Forwarding Rate

Definition: The maximum rate at which the received frames are forwarded by the forwarding function.

Measurement units: Output frames per second at a frame size of N bytes, OR output bits (bytes) per second.

Latency

Definition: For store and forward devices, the time interval starting when the input frame reaches the input port and ending when the output frame is seen on the output port.

For bit forwarding devices (cut-through), the time interval starting when the end of the first bit of the input frame reaches the input port and ending when the start of the first bit of the output frame is seen on the output port.

Measurement units: seconds

Loss Rate

Definition: Percentage of frames that should have been forwarded by the forwarding functions but were not.

Measurement units: Percentage of N byte input frames that are dropped.

Throughput

Definition: The maximum rate at which none of the received frames are dropped by the forwarding function.

Measurement units: Input frames per second at a frame size of N bytes, OR maximum (input or output) bits (bytes) per second.

Maximum LSPs supported at Throughput rate

Definition: The maximum number of LSPs that can be supported at the throughput rate.

Measurement units: number of LSPs in the NHLFE table

4.2 MPLS Terminology

Forwarding Equivalence Class (FEC)

A group of IP packets which are forwarded in the same manner (e.g., over the same path, with the same forwarding treatment).

FEC to Next Hop Label Forwarding Entry Map (FTN)

The "FEC-to-NHLFE" (FTN) maps each FEC to a set of NHLFEs. It is used when forwarding packets that arrive unlabeled, but which are to be labeled before being forwarded.

Incoming Label Map (ILM)

The "Incoming Label Map" (ILM) maps each incoming label to a set of NHLFEs. It is used when forwarding packets that arrive as labeled packets.

Label

A short fixed length physically contiguous identifier, which is used to identify a FEC, usually of local significance. A label is carried within a larger shim header or a Label Stack Entry on media such as Ethernet [10]. On broadcast media such as Ethernet, a shim header is 32-bits – a 20-bit label value field, an 8-bit TTL, 3-bit EXP (Experimental) field, and a 1-bit BoS (Bottom of Stack) flag.

Label Swap

The basic forwarding operation - it performs a lookup on the top incoming label on the label stack to determine the outgoing label, encapsulation, port, and other data handling information.

Label Swapping

A forwarding paradigm allowing streamlined forwarding of data by using labels to identify classes of data packets which are treated indistinguishably when forwarding.

Label Switched Hop

The hop between two MPLS nodes, on which forwarding is done using labels.

Label Switched Path (LSP)

The path through one or more LSRs at one level of the hierarchy followed by a packet in a particular FEC.

Label Switching Router (LSR)

An MPLS node capable of forwarding IPv4 and labeled packets.

Labeled Edge Router (LER)

An MPLS node at the edge of an MPLS network, capable of performing all or some of the following operations:

- Converting an incoming IPv4 packet into an MPLS packet for switching through the MPLS domain.
- Converting an MPLS packet into an IPv4 packet and performing IPv4 forwarding at the egress of an MPLS domain.

The functionality of an LER and LSR may be incorporated into a single box.

Label Stack

A set of labels attached to a packet that must be forwarded through multiple MPLS domains. The topmost label specifies the LSP for forwarding the packet through the current domain.

MPLS Domain

A contiguous set of nodes which forward packets using MPLS and which are also in one Routing or Administrative Domain

MPLS Label

A label which is carried in a packet header, and which represents the packet's FEC

MPLS Node

A node running MPLS. An MPLS node will be aware of MPLS control protocols, may operate one or more layer 3 routing protocols, and will be capable of forwarding packets based on labels. An MPLS node must be capable of forwarding IPv4 packets.

Next Hop Label Forwarding Entry (NHLFE)

The "Next Hop Label Forwarding Entry" (NHLFE) is used when forwarding a labeled packet. It contains the following information:

1. the packet's next hop
2. the operation to perform on the packet's label stack; this is one of the following operations:
 - a. swap the label at the top of the label stack with a specified new label
 - b. pop the topmost label from the label stack
 - c. swap the label at the top of the label stack with a specified new label, and then push one or more specified new labels onto the label stack.
 - d. push one or more labels on the empty label stack.
3. Outgoing label(s) used for label stack operations

It may also contain:

4. the data link encapsulation to use when transmitting the packet
5. the way to encode the label stack when transmitting the packet
6. any other information needed in order to properly dispose of the packet.

Penultimate Hop Popping

The label stack may be popped at the penultimate LSR of the LSP, rather than at the LSP Egress. This saves the egress LSR from doing 2 lookups (MPLS lookup and IPv4 lookup).

5 MPLS Overview

In traditional IPv4 forwarding, a packet jumps from one router to the next, each router making an independent forwarding decision for that packet. Each router chooses the next hop for a packet based on the packet's layer 3 header.

In MPLS forwarding, the forwarding decision is made on the basis of a label in the packet instead of the network header. MPLS integrates a label swapping forwarding paradigm with the network layer routing and thus improves the price/performance of the network layer routing, the scalability of the network layer, and facilitates traffic engineering through an IP network. MPLS techniques are applicable to any network layer protocol and any underlying link layer protocol. Moreover, the MPLS-based approach decouples packet forwarding from the routing, allowing the provision of varied routing services independent of the packet-forwarding paradigm.

An MPLS domain consists of two or more Label Edge Routers (LERs) connected by multiple Label Switched Routers (LSR). Figure 1 shows an example MPLS domain consisting of three LERs connected by four LSRs.

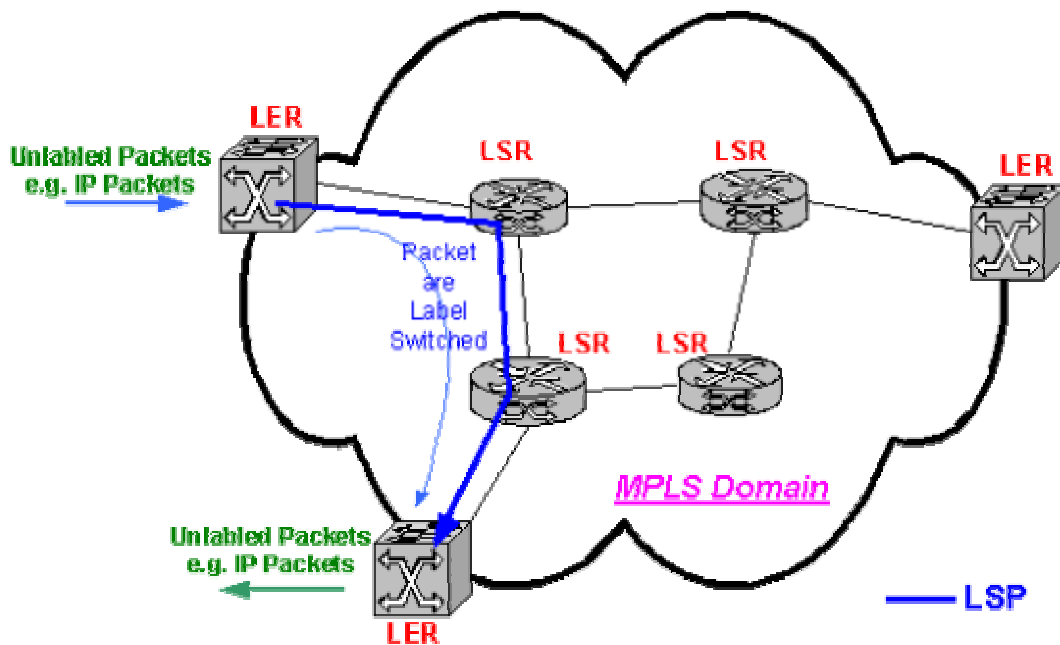


Figure 1 MPLS Domain

Label Switch Paths (LSPs) are setup between ingress and egress LER pairs to transfer packets across the MPLS domain. As such an LSP consists of an ingress LER, one or more LSRs and an egress LER. Multiple LSPs can be setup between any ingress and egress LER pairs. Packets that traverse an MPLS domain undergo varying processing depending upon the location in the LSP where the packet is being processed.

For every packet entering the MPLS domain, ingress LER determines which packets should enter a particular LSP (hence the MPLS domain) and then pushes one or more labels in the packet's label stack. LSRs swap or pop existing labels in the label stack or push one or more new labels on the packet's label stack. Egress LERs are responsible for terminating one or more LSPs by popping the corresponding label(s) from the packet's label stack. It is quite possible that an egress LER may pop all the labels from the packet's label stack, in which case the original packet (e.g. the IP packet) is obtained. Lastly, if multiple MPLS domains are nested, a packet's label stack contains a label for each nested MPLS domain.

Control protocols such as LDP, CR-LDP, or RSVP-TE are used to establish, maintain, and terminate LSPs in an MPLS domain. These control protocols allocate, distribute, assign, release, and withdraw

labels used to realize the LSPs. The control protocols also provision the establishment of constraint-based traffic engineered LSPs called CR-LSP. The constraints could be resource or path requirements through the MPLS domain. CR-LDP is a variant of LDP and is used to establish the CR-LSPs. In order to create LSPs, ingress LER, LSR and egress LER contain tables that need to be populated with control information related to the LSPs being created.

Each ingress LER contains a FEC To NHLFE (FTN) table. A FEC (Forwarding Equivalence Class) is used by the ingress LER to direct packets onto the appropriate LSP. A FEC typically consists of an IP network prefix, an IP host address or an IP five-tuple. Based on this classification, the ingress LER determines the NHLFE (Next Hop Label Forwarding Entry) to use. The NHLFE determines the packets nexthop and the label operation to perform thereby placing it in an LSP.

LSRs and egress LERs also contain an ILM (Incoming Label Map) table, to map the topmost label on an incoming packet's label stack to an NHLFE. The NHLFE contains the operation(s) that must be performed on the packet's label stack. A single MPLS node will usually function as both an LER and an LSR. Hence a node can contain and actively use both the LER and LSR specific tables.

6 Test Configuration

6.1 Reference Design

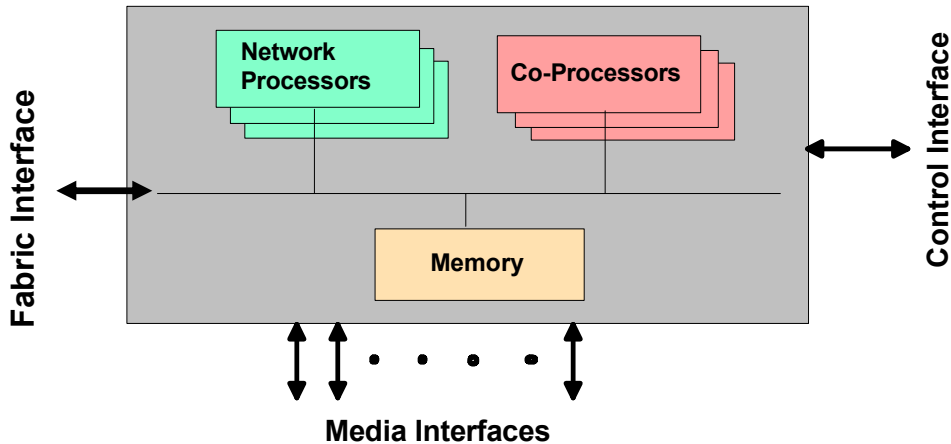


Figure 2 Reference design for MPLS Function Level Benchmarks

The reference design used for the MPLS function level benchmarks is shown in Figure 2. It is assumed to consist of one or more media interfaces and one control interface. The reference design may include multiple network processors and any number of co-processors connected to the network processor in any way. This specification does not define the speed and type of the media and control interfaces. The choice of the media and control interfaces is left to network processor vendors or customers comparing different network processors. When reporting the results of the tests specified in this benchmark, it is required that the reference design is specified in terms of the following parameters in addition to the parameters listed in the Framework document [1]. This reference design is considered to be the device under test (DUT).

REQUIRED information

Block Diagram	Configuration block diagram of the reference design. This diagram should include details on DUT to Traffic Tester connections
Component List	List of hardware components used on the reference design. Should primarily include the NPs, CPs, memory chips, PHYs, framers and fabric interface chips.
Mechanical Size	Size of the base PCB and feature card/daughter card PCB of the reference design
Media Interface(s)	For example: 10/100 Ethernet Gigabit Ethernet OC-3 POS OC-12 POS

	OC-48 POS OC-192 POS etc.
Fabric Interface	List of all fabric interface types on the DUT. For example: CSIX L1, Gigabit Ethernet, ATM, etc.
Number of ports	Total number of media interfaces supported in this reference design. Can be one of the above media interfaces, or a combination of a few of the above. The list should specify the number of interfaces present for each interface type.
Level of Channelization	For example, an OC-48 Network Processor may be capable of processing OC-48c; or capable of OC-48c, 4xOC-12, 16xOC-3, 48xSTS-1 channelization
Network Processor details	Type (part number) and number of NPs used, showing arrangement between NPs: paralleled, pipelined or compound.
Coprocessor details	Type (part number) and number of CPs used, showing arrangement between coprocessors: paralleled, pipelined or compound.
External Memory details	External Memory interfaces – for example SSRAM (what type) , SDRAM (what type) Function of each External memory interface Amount of memory on the reference design for specific functions. Total Control memory used for MPLS tables and IPv4 forwarding tables. Total data memory used for packet buffers, etc.
Control Interface details	PCI/Power PC etc Bandwidth of the interface Relate to some control path benchmarks if bandwidth is sufficient for those functions
Total power consumption	Total power consumption of the reference design at idle condition. Should be accurate within $\pm 10\%$. Idle condition is defined as running the Ingress Traffic Test to measure forwarding rate with no traffic.

DESIRABLE information

Schematic	Reference Design schematics in pdf format
Data Sheets	Data sheets for all the NPs and CPs used in the reference design.
Power consumption per NP/CP	Power consumption breakdown for each NP and CP used in the reference design under the following idle condition. Idle condition is defined as running the IP forwarding application with no traffic.

6.2 Test Setup

The test setup used to obtain the results of the tests described in this document is shown in Figure 3. The media interfaces of the Device Under Test (DUT – same as the reference design shown in Figure 2) are connected to the media interfaces on a data plane tester. It is assumed that the data plane tester is capable of sending and receiving MPLS traffic with different label values and varying label stacks. The data plane tester should also be capable of

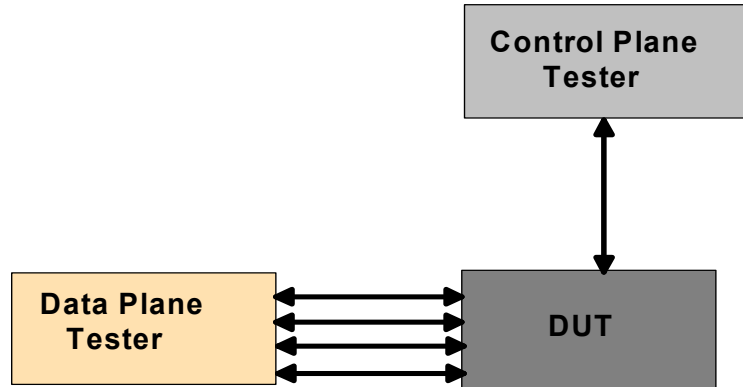


Figure 3 Test setup for MPLS Function Level Benchmarks

calculating the number of packets transmitted and received, and should have the ability to measure the end-to-end latency through the DUT. There are several testers available in the market today that can be used as data plane testers.

The control interface on the DUT is connected to a control plane tester. It is assumed that the control plane tester is capable of generating route update requests, label update and other control messages required for MPLS forwarding.

7 Benchmark Tests

This section describes the MPLS Benchmark tests. The following three configurations are used to benchmark data plane performance of an MPLS implementation.

- Ingress Traffic Test Configuration
- Transit Traffic Test Configuration
- Egress Traffic Test Configuration

Each configuration lists the test parameter values for which the tests must be conducted, followed by the tests used to benchmark the MPLS implementation. All tests **MUST** be performed on each of the parameter values unless specified otherwise in the document. For example, the test to measure the Forwarding rate (7.1.1.2) must be performed with a label stack depth of 1, 2, and 3 labels (7.1.1.1) in the Ingress Traffic Test Configuration.

Optional tests and parameter values are denoted by an asterisk (marked by a ‘*’ sign). If a DUT is unable to perform any of the required tests for any of the required test parameters, a result of “0” must be reported for the test performed with that parameter value. Using the above example, suppose a DUT is unable to measure Forwarding rate for label stack depth of 2 and 3. Then it must perform and report the test for Forwarding rate with a label stack of 1, and report a “0” for Forwarding rate with label stacks of 2 and 3.

7.1 Data Plane Tests

7.1.1 Ingress Traffic Tests

The series of tests described in this section are specifically designed to gauge the performance of the DUT while forwarding IPv4 traffic into an MPLS domain.

7.1.1.1 Ingress Traffic Tests Parameters

This section describes the parameters to be used for the ingress traffic tests. All the tests in section 7.1.1 **MUST** be performed with these parameters.

1. Frame Sizes

The following IP packet sizes (in bytes) must be used for this test.

- IP packet sizes: 40, 110, 238, 494, 1006, 1264, 1500

The link layer frame sizes will be computed from these IP packet sizes based on the media interface. For example, a 40 byte IP packet translates to a 58 byte Ethernet frame (14 byte header + 4 byte CRC). However, the smallest Ethernet frame is 64 bytes. Similarly, a 576 byte IP packet translates to 594 Ethernet frame.

The size of a packet must be selected to be the maximum packet size so that the corresponding link-level frame size never exceeds 1518-bytes when a frame is going into or coming out of a DUT. For example, while testing the operation to PUSH a stack of 3 labels onto a packet, the test for 1500-byte IP packet must be performed with a 1488 byte IP packet for an Ethernet interface. The DUT will push 12 additional bytes onto this packet, and add an 18 byte Ethernet header bringing the outgoing frame size to 1518.

The following streams will be assembled for benchmarking an MPLS implementation.

- A fixed-sized stream of packets for each of the above packet sizes.
- A mixed stream of the below specified packet sizes. The choice of IPv4 packet sizes in this table is based on real world analysis of IP packet size distributions [8]. The packets should be distributed as specified in the last column of the above table. So, 56% of the packets in this stream are 40-byte IP packets, 20% 576-byte packets, and the rest 1500-byte packets.

IP	Ethernet	ATM	POS	Probability
40	64	48	48	0.56
576	594	584	584	0.2
1500	1,518	1,508	1,508	0.24

2. Routing Tables

The following routing tables must be used for this test:

- Mae-West routing table. A snapshot taken from [5] has been standardized for this benchmark.

3. Traffic

The following traffic patterns must be used for this test. Each traffic pattern should be reported as a separate run of the test. The tester should generate traffic on the ports with IP destination addresses drawn randomly from the routing table selected above. The traffic should be uniformly distributed over the ports.

- All frames MUST consist of IPv4 datagrams with no options. All the traffic MUST enter LSPs that push a single label.
- All frames MUST consist of IPv4 datagrams with no options. All the traffic MUST enter LSPs that push 2 labels.
- All frames MUST consist of IPv4 datagrams with no options. All the traffic MUST enter LSPs that push 3 labels.

4. Label Stack Depth

- The packets sent to the DUT MUST have a label stack depth of 0.

5. Label Stack Operation

- An operation to PUSH a single label.
- An operation to PUSH a stack of 2 labels.
- An operation to PUSH a stack of 3 labels.

6. FEC Types

The following FEC type MUST be used for this test:

- IP Prefix classification

7. Test setup

The IPv4 and MPLS forwarding tables should be initialized prior to the test. The number of FECs defined must be at least 200 per media interface on the DUT. The IPv4 forwarding entries should be divided into the defined FECs with each FEC comprising of an equal number of forwarding entries. The NHLFE table should be set up with at least

200 entries per media interface, with a distinct label(s) for each entry. The nexthops in the NHLFE table should be evenly distributed over the media interfaces.

The FTN should be initialized so that the mapping of the FECs to the NHLFEs is one-to-one, each NHLFE being assigned to an FEC.

A separate NHLFE table will be used for each of the traffic patterns described in the above section.

7.1.1.2 Forwarding Rate

Objective: To determine the DUT maximum forwarding rate.

Procedure: Send a specific number of IPv4 packets to the DUT at line rate and count the number of frames received. The maximum forwarding rate is the rate at which frames are received at the data plane tester measured over the time interval required to receive the specific number of frames. Each run of this test should last at least 120 seconds.

Reporting format: The results of this test MUST be reported in the form of a graph. The x coordinate MUST be the frame size in bytes, and the y coordinate MUST be the forwarding rate in frames per second OR bits per second. There MUST be at least 6 lines on the graph. The first 3 lines MUST show the theoretical forwarding rate calculated at various labeled frame sizes for single and multiple label stack operations. The other 3 lines MUST show the measured forwarding rate obtained from the test for single and multiple label stack operations. The results MUST report the aggregate forwarding rate measured over all the media interfaces.

7.1.1.3 Throughput

Objective: To determine the DUT throughput.

Procedure: Send a specific number of frames at a specific rate on all of the media interfaces and count the number of frames received. If the count of the received frames is equal to the count of the transmitted frames, then the throughput is the maximum rate at which frames are received at the tester measured over the time interval required to transmit the specific number of frames. If a fewer number of frames is received than were transmitted, then the rate of transmission is reduced and the test rerun. Each run of the test should last for at least 120 seconds.

Reporting format: Same as in 7.1.1.2. The throughput calculation must include any padding bytes added to a frame as part of the test, as well any inter-frame gap (in bytes) imposed by link-level layers (e.g., 10/100 Ethernet requires a 20-byte gap between successive frames). There is ambiguity in measurement when throughput rate is measured in bits per second, since MPLS packets may cause expansion or contraction of the frame. Therefore, as a convention, the maximum bits per second (either incoming or outgoing rate for the DUT) is reported as the throughput rate. For this test configuration, the rate going out of the DUT MUST be reported as the throughput rate.

7.1.1.4 Latency

Objective: To determine the latency of the DUT.

Procedure: Send frames at a particular frame size through the DUT at a specific rate. Traffic should be run for at least 120 seconds. Each frame should have an identifying tag on it, which is implementation dependent. The time at which the frame is fully transmitted is recorded (timestamp A). The receiver logic

in the data plane tester must recognize the tag information in the frame stream and record the time at which the tagged frame was received (timestamp B). The latency is timestamp B minus timestamp A. The latency MUST be measured for all the frames transmitted as part of this test. The latency test must be repeated for frame rates of 90 and 100% of the throughput rate determined in Section 7.1.1.3.

Reporting format: The latency is reported in the form of graphs, one graph per label stack operation for each prescribed fraction of the throughput rate. The x coordinate MUST be the frame size in bytes and the y coordinate MUST be the measured latency. Each graph MUST have 3 lines: one for the average latency, one for the minimum latency and one for maximum latency at each frame size. The results should report the latencies measured in an aggregate fashion over all the media interfaces.

7.1.1.5 Loss Rate

Objective: To determine the frame loss rate of the DUT over the entire range of input data rates and frame sizes.

Procedure: Send a specific number of frames at a specific rate through the interfaces of the DUT and count the number of frames received at the tester. Traffic should be run for at least 120 seconds. The frame loss rate is calculated using the following formula:

$$(\text{frames sent} - \text{frames received}) * 100 / \text{frames sent}$$

In the first run, the rate is set to 100% media rate. Subsequent trials should be run by reducing the frame rate by 10% increments until there are two successive trials in which no frames are lost.

Reporting format: The Loss rate is reported in the form of 2 graphs where the x coordinate MUST be the transmitted frame rate and the y coordinate MUST be the loss rate for multiple and single label stack operations. Multiple lines on the graph should depict loss rates at different frame sizes.

7.1.1.6 Maximum LSPs supported at throughput rate (*)

Objective: To determine the maximum number of LSPs (the maximum number of entries in the NHLFE table) that can be supported at the throughput rate.

Procedure: Send frames through the DUT at the forwarding rate determined by the throughput test in 7.1.1.3. Increase the number of LSPs in the NHLFE in fixed increments, uniformly across all interfaces and then send packets out on these LSPs until frames are no longer being forwarded at the throughput rate. Note that there must be a mechanism to synchronize between adding LSPs and sending traffic out on them.

Setup: Same as in 7.1.1.1

Reporting format: The maximum number of LSPs at line rate is reported in the form of a graph. The x co-ordinate is the packet sizes, and the y-coordinate is the maximum number of LSPs for each packet size. The number of lines on the graph MUST be 3; one each for the single and multiple label stack operations.

7.1.2 Transit Traffic Tests

The series of tests described in this section are specifically designed to gauge the performance of the DUT when forwarding labeled traffic into an LSP.

7.1.2.1 Transit Traffic Tests Parameters

This section describes the parameters to be used for the transit traffic tests. All the tests in section 7.1.2 MUST be performed with these parameters.

1. Frame Sizes

Same as in 7.1.1.1. The size of a packet must be selected to be the maximum packet size so that the corresponding link-level frame size never exceeds 1518-bytes when a frame is coming out of a DUT. For example, while testing the SWAP-PUSH label operation, the test for the 1500-byte IP packet must be performed with a 1496 byte IP packet for an Ethernet interface. The tester will add an additional 4 byte label before sending the packet into the DUT. The DUT will SWAP this label, PUSH another 4 byte label, and add the 18-byte Ethernet header before transmitting the frame.

The frame size MUST be IP packet size plus L2 Frame header size plus MPLS shim header(s) size.

2. Traffic

The following traffic patterns MUST be used for this test. Each traffic pattern should be reported as a separate run of the test. The tester should generate traffic on all the media ports with labels drawn randomly from the labels assigned in the ILM. The traffic should be uniformly distributed over all the ports.

- All frames should consist of labeled traffic with a single label. The traffic MUST hit the NHLFE entries that swap a single label.
- All frames should consist of labeled traffic with 1 label. The traffic MUST hit the NHLFE entries that swap a label and push an additional label onto the label stack.
- All frames should consist of labeled traffic with 1 label. The traffic MUST hit the NHLFE entries that swap a label and push two additional labels onto the label stack.

3. Label Stack Depth

- The packets sent to the DUT MUST have a label stack depth of 1.

4. Label Stack Operation

- A single label SWAP operation.
- A single label SWAP operation and a label PUSH operation MUST be performed.
- A single label SWAP operation and a 2 label PUSH operation MUST be performed.

5. Test setup

The NHLFE table should be set up with at least 200 entries per media interface with a distinct label for each entry. The nexthops in the NHLFE should be evenly distributed over all the media interfaces. The ILM is initialized to contain at least 200 entries per interface.

The ILM is setup so that there is a one-to-one mapping between an ILM entry and an NHLFE. A separate NHLFE table will be used to benchmark performance for each of the traffic patterns listed.

7.1.2.2 Forwarding Rate

Objective: To determine the DUT maximum forwarding rate.

Procedure: Same as in 7.1.1.2

Reporting format: Same as in 7.1.1.2

7.1.2.3 Throughput

Objective: To determine the DUT throughput.

Procedure: Same as in 7.1.1.3

Reporting format: Same as in 7.1.1.3

7.1.2.4 Latency

Objective: To determine the latency of the DUT.

Procedure: Same as in 7.1.1.4

Reporting format: Same as in 7.1.1.4

7.1.2.5 Loss Rate

Objective: To determine the frame loss rate of the DUT over the entire range of input data rates and frame sizes.

Procedure: Same as in 7.1.1.5

Reporting format: Same as in 7.1.1.5

7.1.2.6 Maximum LSPs supported at throughput rate (*)

Objective: To determine the maximum number of LSPs (the maximum number of entries in the NHLFE table) that can be supported at the throughput rate.

Procedure: Send frames through the DUT at the forwarding rate determined by the throughput test in 7.1.2.3. Increase the number of LSPs in the NHLFE in fixed increments, uniformly across all interfaces and then send packets out on these LSPs until frames are no longer being forwarded at the throughput rate. Note that there must be a mechanism to synchronize between adding LSPs and sending traffic out on them.

Setup: Same as in 7.1.1.6

Reporting format: Same as in 7.1.1.6

7.1.3 Egress Traffic Tests

The series of tests described in this section are specifically designed to gauge the performance of the DUT when forwarding labeled traffic into a domain that does not support MPLS.

Vendors who support Penultimate Hop Popping (PHP) SHOULD setup their egress nodes to perform PHP for these tests. The DUT MUST implement the complete MPLS operation for popping the label. For example, the DUT MUST examine the Bottom-of-Stack (BOS) bit for every MPLS label that is popped in the course of running the benchmark.

7.1.3.1 Egress Traffic Tests Parameters

This section describes the parameters to be used for the egress traffic tests. All the tests in Section 7.1.3 MUST be performed with these parameters.

1. Frame Sizes

Same as in 7.1.2.1

2. Routing Tables

The following routing tables must be used for this test (Only vendors not supporting PHP need to use this table):

- Mae-West routing table. A snapshot taken from [5] has been standardized for this benchmark.

3. Traffic

The following traffic patterns must be used for this test. Each traffic pattern should be reported as a separate run of the test. The tester should generate traffic on the ingress ports with labels drawn randomly from the labels in the ILM. The IPv4 packet encapsulated should have destination addresses drawn randomly from the routing table selected above. The traffic should be uniformly distributed over the ports.

- All frames should consist of labeled traffic with a single label.
- All frames should consist of traffic labeled with two labels.
- All frames should consist of traffic labeled with three labels.

4. Label Stack Depth

- Packets sent to the DUT MUST have label stack depth of 1, 2, and 3 depending on the label stack operation.

5. Label Stack Operation

- A single label POP operation MUST be performed.
- An operation to POP all labels in the label stack MUST be performed.

6. Test setup

For vendors not supporting PHP, the IPv4 and MPLS forwarding tables should be initialized prior to the test. IPv4 forwarding tables should be initialized with the entries in the routing tables under test. The next hops for the forwarding table entries should be uniformly distributed over all media interfaces. The ILM is initialized to contain at least 200 entries per media interface.

A separate NHLFE table is used to benchmark performance for each of the traffic patterns listed above.

The ILM entries are uniformly mapped to the NHLFE Table.

7.1.3.2 Forwarding Rate

Objective: To determine the DUT maximum forwarding rate.

Procedure: Same as in 7.1.1.2

Reporting format: Same as in 7.1.1.2

7.1.3.3 Throughput

Objective: To determine the DUT throughput.

Procedure: Same as in 7.1.1.3

Reporting format: Same as in 7.1.1.3, except that the rate coming into the DUT MUST be reported as the throughput rate.

7.1.3.4 Latency

Objective: To determine the latency of the DUT.

Procedure: Same as in 7.1.1.4

Reporting format: Same as in 7.1.1.4

7.1.3.5 Loss Rate

Objective: To determine the frame loss rate of the DUT over the entire range of input data rates and frame sizes.

Procedure: Same as in 7.1.1.5

Reporting format: Same as in 7.1.1.5

APPENDIX A INFORMATIVE ANNEXES

This section contains the informative and compulsory annex to this MPLS Forwarding Benchmark Specification. It provides descriptions associated with the benchmark routing tables, overviews the traffic streams required to run the benchmark tests, and reviews a reference implementation of the benchmark.

A.1. TRAFFIC GENERATION

A.1.1. INGRESS TRAFFIC TESTS

This section describes how the Mae West route table should be used for generating traffic to perform the Ingress Traffic Tests, specified in Section 7.1.1, with a particular port configuration.

The Mae West snapshot [12] must be used to perform all the Ingress Traffic Tests for all label stack operations mandated by Section 7.1.1. This involves generating the routing table entries for the data traffic.

An IPv4 Prefix Classifier will classify each packet based on the destination IP address of the packet. This IP address is obtained from a snapshot of the Mae West table [12]. The parseMaeWest.tcl script [13] must be used to select the Mae West entries for the data traffic to hit. This script is a generic TCL script independent of any particular data plane tester. The script input is the Mae West table, the desired number of route subsets, and the desired number of entries in each subset. The algorithm is deterministic and selects route table entries based on the prefix length distribution of the full Mae West table and attempts to provide a similar prefix length distribution in the subset of entries selected. The script output is a text file providing a table of the selected routes, the subset to which they belong, their prefix length and the IP destination addresses. The data plane tester must exercise the IP destination addresses from this list in a sequence of no more than 3 packets per address with each port using a different subset. The traffic sent in must exercise the number of LSPs prescribed by Section 7.1.1.

The entire Mae West snapshot route table **MUST** be loaded into the table of entries for the classifier in the DUT. Similarly, the MPLS tables corresponding to the classified packets must be setup to perform one of the above label stack operations and forward the packet.

A.1.2. TRANSIT TRAFFIC TESTS

This section describes the scheme to generate traffic for the Transit traffic tests specified in Section 7.1.2.

Each packet going into the DUT must have a single label on it. The Incoming Label Map (ILM) entry corresponding to the label value must point to an entry that will perform one of the label operations mandated by Section 7.1.2. The traffic going into the DUT must exercise the number of LSPs prescribed by Section 7.1.2, and all forwarding tables corresponding to these LSPs on the DUT must be populated before the test. A scheme to select label(s) for each LSP is presented in the Implementation Kit [11]. However, no specific script or scheme is required for selecting labels.

A.1.3. EGRESS TRAFFIC TESTS

The traffic generation requirements for Egress traffic tests are similar to the traffic requirements for the Transit traffic tests.

Each packet going into the DUT must contain a stack of MPLS labels as specified in Section 7.1.3. The traffic going into the DUT must exercise the number of LSPs prescribed in Section 7.1.3. All forwarding tables corresponding to these LSPs on the DUT must be populated to perform the label stack operations, specified in Section 7.1.3, before the test. The DUT must be setup such that Penultimate Hop Popping (PHP) is enabled. An IPv4 Lookup must not be performed in the course of forwarding the packet.

A.2. IMPLEMENTATION KIT

An example implementation of the NPF MPLS forwarding application level benchmark tests is provided in the MPLS Benchmark Implementation Kit [11]. This implementation kit is a package provided in a zip file. Please refer to the user manual provided with the zip file for additional details. The Tcl scripts in this toolkit are provided as is. Modifications may be needed in different DUT environments but these scripts should still provide an excellent way to get started.

This implementation kit is provided purely as an example and its use is not mandatory (other than the use of the Mae West snapshot [12] and the parseMaeWest.tcl script). The techniques used to generate traffic for the Ingress, Transit and Egress traffic tests are presented in the User's Manual in the Implementation kit [11].