

Network Working Group  
Internet Draft  
Intended status: Standards Track  
Expires: April 15, 2011

Siva Sivabalan (Ed.)  
Sami Boutros (Ed.)  
Luca Martini  
  
Cisco Systems, Inc.

October 15, 2010

MAC Address Withdrawal over Static Pseudowire  
draft-boutros-pwe3-mpls-tp-mac-wd-00.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 April 15, 2011.

Abstract

This document specifies a mechanism to signal MAC address withdrawal notification using PW Associated Channel (ACH). Such notification is useful when statically provisioned PWs are deployed in VPLS/H-VPLS environment.

This document is a product of a joint Internet Engineering Task Force(IETF) / International Telecommunication Union



## Table of Contents

1. Introduction.....	2
2. Terminology.....	3
3. MAC Withdraw OAM Message.....	3
4. Operation.....	5
4.1.1. Operation of Sender.....	5
4.1.2. Operation of Receiver.....	5
5. Security Considerations.....	5
6. IANA Considerations.....	5
7. References.....	6
7.1. Normative References.....	6
7.2. Informative References.....	6
Author's Addresses.....	7
Full Copyright Statement.....	7
Intellectual Property Statement.....	8

## 1. Introduction

An LDP-based MAC Address Withdrawal Mechanism is specified in RFC4762 [2] to remove dynamically learned MAC addresses when the source of those addresses can no longer forward traffic. This is accomplished by sending an LDP Address Withdraw Message with a MAC List TLV containing the MAC addressed to be removed to all other PEs over LDP sessions. When the number of MAC addresses to be removed is large, empty MAC List TLV may be used. [3] describes an optimized MAC withdrawal mechanism which can be used to remove only the set of MAC addresses that need to be re-learned in H-VPLS networks. The solution also provides optimized MAC Withdrawal operations in PBB-VPLS networks.

A PW can be signaled via LDP or can be statically provisioned. In the case of static PW, LDP based MAC withdrawal mechanism cannot be used. This is analogous to the problem and solution described in [4] where PW OAM message has been introduced to carry PW status TLV using in-band PW Associated Channel. In this document, we propose to use PW OAM message to withdraw MAC address(es) learned via static PW.

Internet-Draft draft-boutros-pwe3-mpls-tp-mac-wd-00.txt October 2010  
This document is a product of a joint Internet Engineering Task Force (IETF) / International Telecommunication Union Telecommunication Standardization Sector (ITU-T) effort to include an MPLS Transport Profile within the IETF MPLS and PWE3 architectures to support the capabilities and functionalities of a packet transport network.

Conventions used in 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 RFC-2119 [1].

## 2. Terminology

ACK: Acknowledgement.

LDP: Label Distribution Protocol.

MAC: Media Access Control

MPLS: Multi Protocol Label Switching.

OAM: MPLS Operations, Administration and Maintenance.

PE: Provide Edge Node.

PW: PseudoWire.

TLV: Type, Length, and Value.

VPLS: Virtual Private LAN Services.

## 3. MAC Withdraw OAM Message

LDP provides a reliable packet transport for control plackets for dynamic PWs. This can be contrasted with static PWs which rely on re-transmission and acknowledgments (ACK) for reliable OAM packet delivery as described in [4]. The proposed solution for MAC withdrawal over static PW also relies on re-transmissions and ACKs. However, ACK is mandatory. A given MAC withdrawal notification is sent as a PW OAM message, and the sender keeps re-transmitting the message until it receives an ACK for that message. Once a receiver successfully remove MAC address(es) in response to a MAC address withdraw OAM message, it should not unnecessarily remove MAC address(es) upon getting refresh message(s). To facilitate this, the

The format of the MAC address withdraw OAM message is shown in Figure 1. The PW OAM message header is exactly the same as what is defined in [4]. Since the MAC withdrawal PW OAM message is not refreshed for ever, the "Refresh Timer" field in the message header is not used. A MAC address withdraw OAM message MUST contain a "Sequence Number TLV" otherwise the entire message is dropped. It may contain PE-ID or MAC Flush Parameter TLVs defined in [3] when static PWs are deployed in H-VPLS and PBB-VPLS scenarios. The sequence number TLV has U (Unknown) and F (Forward) bits set to 1 and 0 respectively so that if a receiver does not recognize the TLV, it drops the whole message.

In this section, MAC List TLV, PE-ID TLV, and MAC Flush Parameter TLV are collectively referred to as "MAC TLV(s)". The processing rules of MAC List TLV is governed by [2], and the corresponding rules of PE-ID TLV and MAC Flush Parameter TLV are governed by [3].

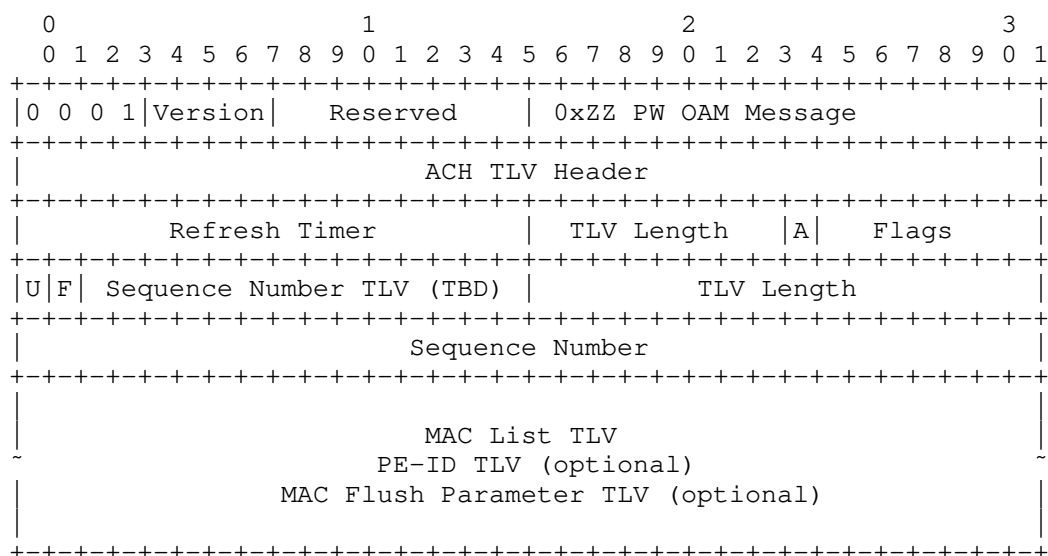


Figure 1: MAC Address Withdraw PW OAM Packet Format.

An ACK for MAC withdraw OAM message is the same as the one shown in Figure 1 except that:

- . A-bit is set.
- . It does not include MAC TLV(s).

#### 4. Operation

This section describes how the initial MAC withdraw OAM messages are sent and retransmitted, as well as how the messages are processed and retransmitted messages are identified.

##### 4.1.1. Operation of Sender

Each PW is associated with a counter to keep track of the sequence number of the transmitted MAC withdrawal messages. Whenever a node sends a new set of MAC TLVs, it increments the transmitted sequence number counter, and include the new sequence number in the message.

The sender expects an ACK from the receiver within a time interval which we call "Retransmit Time" which can be either a default or configured value. If the ACK arrives within the Retransmit Time, the sender assumes that the message transmission is successful. Otherwise, it retransmits the message with the same sequence number as the original message.

##### 4.1.2. Operation of Receiver

Each PW is associated with a counter to keep track of the sequence number of the MAC withdrawal message received last. Whenever a MAC withdrawal message is received, and if the sequence number on the message is greater than the receive counter, the MAC address(es) contained in the MAC TLV(s) is/are removed, and the receive counter is incremented. The receiver sends an ACK whose sequence number is the same as the received message.

If the sequence number in the received message is smaller than or equal to the receive counter, the MAC TLV(s) is/are not processed. However, an ACK whose sequence number is the same as the received message is sent.

#### 5. Security Considerations

This document does not introduce any additional security constraints.

#### 6. IANA Considerations

IANA needs to assign the type value for Sequence Number TLV.

## 7. References

### 7.1. Normative References

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

### 7.2. Informative References

- [2] Mark Lassere, et. al, "Virtual Private LAN Service (VPLS) Using Label Distribution Protocol (LDP) Signaling", RFC4762, January 2007.
- [3] Pranjal Kumar Dutta, et. al, "LDP Extensions for Optimized MAC Address Withdrawal in H-VPLS", draft-ietf-l2vpn-vpls-ldp-mac-opt-02.txt (work in progress), July 2010.
- [4] Luca Martini, et. al, "Pseudowire Status for Static Pseudowires", draft-ietf-pwe3-static-pw-status-00.txt (work in progress), February 2010.

Siva Sivabalan  
Cisco Systems, Inc.  
2000 Innovation Drive  
Kanata, Ontario, K2K 3E8  
Canada  
Email: msiva@cisco.com

Sami Boutros  
Cisco Systems, Inc.  
3750 Cisco Way  
San Jose, California 95134  
USA  
Email: sboutros@cisco.com

Luca Martini  
Cisco Systems, Inc.  
9155 East Nichols Avenue, Suite 400  
Englewood, CO, 80112  
United States  
Email: lmartini@cisco.com

#### Full Copyright Statement

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.

All IETF Documents and the information contained therein are provided on an "AS IS" basis and THE CONTRIBUTOR, THE ORGANIZATION HE/SHE REPRESENTS OR IS SPONSORED BY (IF ANY), THE INTERNET SOCIETY, THE IETF TRUST AND THE INTERNET ENGINEERING TASK FORCE DISCLAIM ALL WARRANTIES, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO ANY WARRANTY THAT THE USE OF THE INFORMATION THEREIN WILL NOT INFRINGE

## Intellectual Property Statement

The IETF takes no position regarding the validity or scope of any Intellectual Property Rights or other rights that might be claimed to pertain to the implementation or use of the technology described in this document or the extent to which any license under such rights might or might not be available; nor does it represent that it has made any independent effort to identify any such rights.

Copies of Intellectual Property disclosures made to the IETF Secretariat and any assurances of licenses to be made available, or the result of an attempt made to obtain a general license or permission for the use of such proprietary rights by implementers or users of this specification can be obtained from the IETF on-line IPR repository at <http://www.ietf.org/ipr>.

The IETF invites any interested party to bring to its attention any copyrights, patents or patent applications, or other proprietary rights that may cover technology that may be required to implement any standard or specification contained in an IETF Document. Please address the information to the IETF at [ietf-ipr@ietf.org](mailto:ietf-ipr@ietf.org).

The definitive version of an IETF Document is that published by, or under the auspices of, the IETF. Versions of IETF Documents that are published by third parties, including those that are translated into other languages, should not be considered to be definitive versions of IETF Documents. The definitive version of these Legal Provisions is that published by, or under the auspices of, the IETF. Versions of these Legal Provisions that are published by third parties, including those that are translated into other languages, should not be considered to be definitive versions of these Legal Provisions.

For the avoidance of doubt, each Contributor to the IETF Standards Process licenses each Contribution that he or she makes as part of the IETF Standards Process to the IETF Trust pursuant to the provisions of RFC 5378. No language to the contrary, or terms, conditions or rights that differ from or are inconsistent with the rights and licenses granted under RFC 5378, shall have any effect and shall be null and void, whether published or posted by such Contributor, or included with or in such Contribution.

Internet-Draft draft-boutros-pwe3-mpls-tp-mac-wd-00.txt October 2010  
Acknowledgment

Funding for the RFC Editor function is currently provided by the  
Internet Society.

Internet Engineering Task Force  
Internet-Draft  
Intended status: Standards Track  
Expires: April 21, 2011

G. Chen  
L. Li  
China Mobile  
October 18, 2010

IPv6 Provider Edge Routers (6PE) Information Base (MIB)  
draft-chen-mpls-6pe-mib-00

Abstract

This memo defines a portion of the Management Information Base (MIB) for use with network management protocols in the Internet community. In particular, it describes a MIB module for IPv6 Provider Edge Routers (6PE) over Multiprotocol Label Switching (MPLS) Label Switching Routers (LSRs).

Status of this Memo

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

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

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

This Internet-Draft will expire on April 15, 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.

#### Table of Contents

1. Introduction . . . . .	3
2. The Internet-Standard Management Framework . . . . .	3
3. Overview of MIB objects . . . . .	3
3.1. 6PETunnelIfTable . . . . .	4
3.2. 6PEMplsIfTable . . . . .	4
4. 6PE-MPLS-STD-MIB Module Definitions . . . . .	4
5. Security Considerations . . . . .	7
6. IANA Considerations . . . . .	7
7. Normative References . . . . .	7
Authors' Addresses . . . . .	8

## 1. Introduction

IPv6 Provider Edge Routers (6PE) is a IPv6 transition technology, which could shift network to provide IPv6 access depending on existing Multiprotocol Label Switching (MPLS) core network. Operators could deploy IPv6 network without modifying IPv4 enable MPLS cloud. Therefore, 6PE is treated as a IPv6 transition solution on the early stage. 6PE will be adapted in more and more operational IP networks on account of IPv4 depletion and incremental advantages.

RFC 4789[RFC4789] has elaborated 6PE technology. When tunneling IPv6 packets over the IPv4 MPLS backbone, rather than successively prepend an IPv4 header and then perform label imposition based on the IPv4 header, the ingress 6PE Router MUST directly perform label imposition of the IPv6 header without prepending any IPv4 header. In respect of managing IPv6 tunnel, RFC 4087[RFC4087] has specified managed objects used for managing tunnels of any type over IPv4 and IPv6 networks. But, this MIB module does not support tunnels over non-IP networks. RFC4382[RFC4382] has defined managed objects to configure and monitor MPLS layer 3 Virtual Private Networks. Nevertheless, 6PE is neither Layer 3 IP tunnel nor MPLS layer 3 VPN. This document is aimed at describing managed objects for 6PE.

## 2. The Internet-Standard Management Framework

For a detailed overview of the documents that describe the current Internet-Standard Management Framework, please refer to section 7 of RFC 3410[RFC3410] .

Managed objects are accessed via a virtual information store, termed the Management Information Base or MIB. MIB objects are generally accessed through the Simple Network Management Protocol (SNMP). Objects in the MIB are defined using the mechanisms defined in the Structure of Management Information (SMI). This memo specifies a MIB module that is compliant to the SMIV2, which is described in STD 58, RFC 2578[RFC2578], STD 58, RFC 2579[RFC2579] and STD 58, RFC 2580[RFC2580]

## 3. Overview of MIB objects

The following subsections describe the purpose of each of the objects contained in the 6PE-MPLS-STD-MIB.

### 3.1. 6PETunnelIfTable

6PETunnelIfTable are defined in the MIBs defining the encapsulation. An entry in the 6PE Tunnel MIB will exist for every interface entry with this interface type. An implementation of the 6PE Tunnel MIB may allow 6PETunnelIfTable to be created. Creating a tunnel will also add an entry in the 6PETunnelIfTable, and deleting a tunnel will likewise delete the entry in the 6PETunnelIfTable.

### 3.2. 6PEmplsIfTable

This table controls MPLS-specific parameters when 6PE is going to be carried over MPLS cloud.

## 4. 6PE-MPLS-STD-MIB Module Definitions

IMPORTS

MODULE-IDENTITY, OBJECT-TYPE, transmission,

Integer32, IpAddress FROM SNMPv2-SMI -- [RFC2578]

RowStatus, StorageType FROM SNMPv2-TC -- [RFC2579]

MODULE-COMPLIANCE,

OBJECT-GROUP FROM SNMPv2-CONF -- [RFC2580]

InetAddressType,

InetAddress FROM INET-ADDRESS-MIB -- [RFC4001]

ifIndex,

InterfaceIndexOrZero FROM IF-MIB -- [RFC2863]

MplsTunnelIndex, MplsTunnelInstanceIndex,

MplsLdpIdentifier, MplsLsrIdentifier

FROM MPLS-TC-STD-MIB -- [RFC3811]

MplsIndexType

FROM MPLS-LSR-STD-MIB -- [RFC3813]

```
6peMplsStdMIB MODULE-IDENTITY
LAST-UPDATED "201010180000Z" -- 18 October 2010 00:00:00 GMT
ORGANIZATION "IPv6 Provider Edge Routers (6PE) Working Group."
CONTACT-INFO
"
    Chen Gang, Editor
    Email: chengang@chinamobile.com

    Li Lianyan, Editor
    Email: lilianyan@chinamobile.com
"
DESCRIPTION
    "This MIB module complements the 6PE-MPLS-STD-MIB for 6PE.

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

-- Revision history.
REVISION "201010180000Z" -- 18 October 2010 00:00:00 GMT
DESCRIPTION
    "First published"

 ::= { 6peMplsStdMIB 1 }

-- 6PETunnelIfTable.

6PETunnelIfTable OBJECT-TYPE
SYNTAX      SEQUENCE OF TunnelIfEntry
MAX-ACCESS not-accessible
STATUS      current
DESCRIPTION
    "The (conceptual) table containing information on
    6PE tunnels."
 ::= { 6peMplsStdMIB 1 }

6PETunnelIfEntry OBJECT-TYPE
SYNTAX      TunnelIfEntry
MAX-ACCESS not-accessible
STATUS      current
DESCRIPTION
    "An entry (conceptual row) containing the information
    on a particular configured 6PE tunnel."
INDEX       { ifIndex }
 ::= { 6PETunnelIfTable 1 }

6PETunnelIfEntry ::= SEQUENCE {
    6PETunnelIfHopLimit      Integer32,
```

```
6PETunnelIfSecurity          INTEGER,
6PETunnelIfTOS               Integer32,
6PETunnelIfFlowLabel         IPv6FlowLabelOrAny,
6PETunnelIfLocalAddress      InetAddress,
6PETunnelIfRemoteAddress     InetAddress
}

6PETunnelIfLocalAddress OBJECT-TYPE
SYNTAX      IPAddress
MAX-ACCESS  read-only
STATUS      deprecated
DESCRIPTION
    "The address of the local endpoint of the tunnel"
 ::= { 6PETunnelIfEntry 1 }

6PETunnelIfRemoteAddress OBJECT-TYPE
SYNTAX      IPAddress
MAX-ACCESS  read-only
STATUS      deprecated
DESCRIPTION
    "The address of the remote endpoint of the tunnel"
 ::= { 6PETunnelIfEntry 2 }

6PETunnelIfHopLimit OBJECT-TYPE
SYNTAX      Integer32 (0 | 1..255)
MAX-ACCESS  read-write
STATUS      current
DESCRIPTION
    "The IPv6 Hop Limit to use in IPv6 header."
 ::= { 6PETunnelIfEntry 3 }

6PETunnelIfTOS OBJECT-TYPE
SYNTAX      Integer32 (-2..63)
MAX-ACCESS  read-write
STATUS      current
DESCRIPTION
    "The method used to set IPv6 Traffic Class in IP header."
 ::= { 6PETunnelIfEntry 4 }

        6PETunnelIfFlowLabel OBJECT-TYPE
SYNTAX      IPv6FlowLabelOrAny
MAX-ACCESS  read-write
STATUS      current
DESCRIPTION
    "The method used to set the IPv6 Flow Label value."
 ::= { 6PETunnelIfEntry 5 }
```

```
-- 6PEmplsIfTable.

6PEmplsIfTable    OBJECT-TYPE
SYNTAX            SEQUENCE OF PwMplsEntry
MAX-ACCESS        not-accessible
STATUS            current
DESCRIPTION
    "This table controls MPLS-specific parameters when the 6PE is
    going to be carried over MPLS cloud."
 ::= { 6peMplsStdMIB 2 }

6PEmplsEntry      OBJECT-TYPE
SYNTAX            6PEmplsEntry
MAX-ACCESS        not-accessible
STATUS            current
DESCRIPTION
    "A row in this table represents parameters specific to MPLS
    cloud for 6PE."

INDEX { 6PEIndex }

 ::= { 6PEmplsIfTable 1 }
```

## 5. Security Considerations

It needs to be further identified.

## 6. IANA Considerations

This memo includes no request to IANA.

## 7. Normative References

- [RFC2578] McCloghrie, K., Ed., Perkins, D., Ed., and J. Schoenwaelder, Ed., "Structure of Management Information Version 2 (SMIv2)", STD 58, RFC 2578, April 1999.
- [RFC2579] McCloghrie, K., Ed., Perkins, D., Ed., and J. Schoenwaelder, Ed., "Textual Conventions for SMIv2", STD 58, RFC 2579, April 1999.
- [RFC2580] McCloghrie, K., Perkins, D., and J. Schoenwaelder,

"Conformance Statements for SMIV2", STD 58, RFC 2580, April 1999.

- [RFC2863] McCloghrie, K. and F. Kastenholz, "The Interfaces Group MIB", RFC 2863, June 2000.
- [RFC3410] Case, J., Mundy, R., Partain, D., and B. Stewart, "Introduction and Applicability Statements for Internet-Standard Management Framework", RFC 3410, December 2002.
- [RFC3811] Nadeau, T. and J. Cucchiara, "Definitions of Textual Conventions (TCs) for Multiprotocol Label Switching (MPLS) Management", RFC 3811, June 2004.
- [RFC3813] Srinivasan, C., Viswanathan, A., and T. Nadeau, "Multiprotocol Label Switching (MPLS) Label Switching Router (LSR) Management Information Base (MIB)", RFC 3813, June 2004.
- [RFC4001] Daniele, M., Haberman, B., Routhier, S., and J. Schoenwaelder, "Textual Conventions for Internet Network Addresses", RFC 4001, February 2005.
- [RFC4087] Thaler, D., "IP Tunnel MIB", RFC 4087, June 2005.
- [RFC4382] Nadeau, T. and H. van der Linde, "MPLS/BGP Layer 3 Virtual Private Network (VPN) Management Information Base", RFC 4382, February 2006.
- [RFC4789] Schoenwaelder, J. and T. Jeffree, "Simple Network Management Protocol (SNMP) over IEEE 802 Networks", RFC 4789, November 2006.

#### Authors' Addresses

Gang Chen  
China Mobile  
53A, Xibianmennei Ave.,  
Xuanwu District,  
Beijing 100053  
China

Email: chengang@chinamobile.com

Lianyuan Li  
China Mobile  
53A,Xibianmennei Ave.  
Beijing 100053  
P.R.China

Phone: +86-13910750201  
Email: lilianyuan@chinamobile.com



Network Working Group  
Internet Draft  
Intended status: Informational  
Expires: April 25, 2011

Luyuan Fang  
Cisco Systems  
Scott  
Ben Niven-Jenkins  
Velocix  
Raymond Zhang  
BT  
Nabil Bitar  
Verizon  
Masahiro DAIKOKU  
KDDI  
Scott Mansfield  
Ericsson  
Lei Wang  
Telenor

October 25, 2010

Security Framework for MPLS-TP  
draft-fang-mpls-tp-security-framework-03.txt

Abstract

This document provides a security framework for Multiprotocol Label Switching Transport Profile (MPLS-TP). MPLS-TP Requirements and MPLS-TP Framework are defined in [RFC 5654] and [RFC 5921]. Extended from MPLS technologies, MPLS-TP introduces new OAM capabilities, transport oriented path protection mechanism, and strong emphasis on static provisioning supported by network management systems. This document addresses the security aspects that are relevant in the context of MPLS-TP specifically. It describes the security requirements for MPLS-TP; potential securities threats and migration procedures for MPLS-TP networks and MPLS-TP inter-connection to MPLS, GMPLS networks. The general security analysis and guidelines for MPLS and GMPLS are addressed in [MPLS/GMPLS Security FW], will not be covered in this document.

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.

MPLS-TP Security framework  
October 2010

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 January 12, 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 in effect on the date of publication of this document (<http://trustee.ietf.org/license-info>). Please review these documents carefully, as they describe your rights and restrictions with respect to this document.

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
1.1. Background and Motivation.....	3
1.2. Scope.....	4
1.3. Terminology.....	5
1.4. Structure of the document.....	6
2. Security Reference Models.....	6

2.1. Security Reference Model 1.....	7
2.2. Security Reference Model 2.....	8
3. Security Requirements for MPLS-TP.....	11
3.1. Protection within the MPLS-TP Network.....	11
4. Security Threats.....	13
4.1. Attacks on the Control Plane.....	15
4.2. Attacks on the Data Plane.....	15
5. Defensive Techniques for MPLS-TP Networks.....	16
5.1. Authentication.....	16
5.2. Access Control Techniques.....	17
5.3. Use of Isolated Infrastructure.....	18
5.4. Use of Aggregated Infrastructure.....	18
5.5. Service Provider Quality Control Processes.....	18
5.6. Verification of Connectivity.....	18
6. Monitoring, Detection, and Reporting of Security Attacks.....	18
7. Security Considerations.....	19
8. IANA Considerations.....	19
9. Normative References.....	19
10. Informative References.....	20
11. Author's Addresses.....	20

## Requirements Language

Although this document is not a protocol specification, 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 [RFC 2119].

## 1. Introduction

### 1.1. Background and Motivation

This document provides a security framework for Multiprotocol Label Switching Transport Profile (MPLS-TP).

MPLS-TP Requirements and MPLS-TP Framework are defined in [RFC 5654] and [RFC 5921]. The intent of MPLS-TP development is to address the needs for transport evolution, the fast growing bandwidth demand accelerated by new packet based services and multimedia applications, from Ethernet Services, Layer 2 and Layer 3 VPNS, triple play to Mobile Access Network (RAN) backhaul, etc.

MPLS-TP is based on MPLS technologies to take advantage of the technology maturity, and it is required to maintain the transport characteristics.

Focused on meeting the transport requirements, MPLS-TP uses a subset of MPLS features, and introduces extensions to reflect the transport technology characteristics. The added functionalities include in-band OAM, transport oriented path protection and recovery mechanisms, etc. There is strong emphasis on static provisioning supported by Network Management System (NMS) or Operation Support System (OSS). Of course, there are needs for MPLS-TP and MPLS interworking.

The security aspects for the new extensions which are particularly designed for MPLS-TP need to be addressed. The security models, requirements, threat and defense techniques previously defined in [RFC 5920] can be used for the re-use of the existing functionalities in MPLS and GMPLS, but not sufficient to cover the new extensions.

## 1.2. Scope

This document addresses the security aspects that are specific to MPLS-TP. It intends to provide the security requirements for MPLS-TP; defines security models which apply to various MPLS-TP deployment scenarios; identifies the potential securities threats and migration procedures for MPLS-TP networks and MPLS-TP inter-connection to MPLS, GMPLS networks. Inter-AS and Inter-provider security for MPLS-TP to MPLS-TP connections or MPLS-TP to MPLS connections are discussed, where connections present higher security risk factors than connections for Intra-AS MPLS-TP.

The general security analysis and guidelines for MPLS and GMPLS are addressed in [MPLS/GMPLS Security FW], the content which has no new impact to MPLS-TP will not be repeated in this document. Other general security issues regarding transport networks but not specific to MPLS-TP is out of scope as well. Readers may also refer to the "Security Best Practices Efforts and Documents" [opsec effort] and "Security Mechanisms for the Internet" [RFC3631] (if there are linkage to internet in the applications) for general network operation security considerations. This document does not intend to define the specific mechanisms/methods which must be implemented to satisfy the security requirements.

Issues/Areas to be addressed:

G-Ach (control plane attack, DoS attack, message intercept, etc.)

Spoofing ID  
Loopback  
NMS attack  
NMS and CP interaction  
MIP/MEP assignment and attacks  
Topology discovery  
Data plane authentication  
Label authentication  
DoS attack in Data Plane  
Performance Monitoring

This draft is work in progress.

### 1.3. Terminology

This document uses MPLS, MPLS-TP, and Security specific terminology. Detailed definitions and additional terminology for MPLS-TP may be found in [RFC 5654], [RFC 5921], and MPLS/GMPLS security related terminology in [RFC 5920].

Term	Definition
-----	
APS	Automatic Protection Switching
ATM	Asynchronous Transfer Mode
BFD	Bidirectional Forwarding Detection
CE	Customer-Edge device
CM	Configuration Management
CoS	Class of Service
CPU	Central Processing Unit
DNS	Domain Name System
DoS	Denial of Service
EMF	Equipment Management Function
ESP	Encapsulating Security Payload
FEC	Forwarding Equivalence Class
FM	Fault Management
GAL	Generic Alert Label
G-ACH	Generic Associated Channel
GMPLS	Generalized Multi-Protocol Label Switching
GCM	Galois Counter Mode
IKE	Internet Key Exchange
LDP	Label Distribution Protocol
LMP	Link Management Protocol
LSP	Label Switched Path
MD5	Message Digest Algorithm
MEP	Maintenance End Point

MIP	Maintenance Intermediate Point
MPLS	MultiProtocol Label Switching
NTP	Network Time Protocol
OAM	Operations, Administration, and Management
PE	Provider-Edge device
PM	Performance Management
PSN	Packet-Switched Network
PW	Pseudowire
QoS	Quality of Service
RSVP	Resource Reservation Protocol
RSVP-TE	Resource Reservation Protocol with Traffic Engineering Extensions
SCC	Signaling Communication Channel
SDH	Synchronous Digital Hierarchy
SLA	Service Level Agreement
SNMP	Simple Network Management Protocol
SONET	Synchronous Optical Network
S-PE	Switching Provider Edge
SRLG	Shared Risk Link Group
SSH	Secure Shell
SSL	Secure Sockets Layer
SYN	Synchronize packet in TCP
TCP	Transmission Control Protocol
TDM	Time Division Multiplexing
TE	Traffic Engineering
TLS	Transport Layer Security
TTL	Time-To-Live
T-PE	Terminating Provider Edge
UDP	User Datagram Protocol
VPN	Virtual Private Network
WG	Working Group of IETF
WSS	Web Services Security

#### 1.4. Structure of the document

Section 1: Introduction  
Section 2: MPLS-TP Security Reference Models  
Section 3: Security Requirements  
Section 4: Security threats  
Section 5: Defensive/mitigation techniques/procedures

Note that this document is currently work in progress, not all requirements and security discussions are included, and some sections will be filled in later revision.

## 2. Security Reference Models

This section defines a reference model for security in MPLS-TP networks.

The models are built on the architecture of MPLS-TP defined in [RFC 5921]. The SP boundaries play the important role to determine the security models for any particular deployment.

This document defines the zone where the single SP has the total operational control to be a trusted zone for that SP. A primary concern is about security aspects that relate to breaches of security from the "outside" of a trusted zone to the "inside" of this zone.

## 2.1. Security Reference Model 1

In the reference model 1, a single SP has the total control of PE/T-PE to PE/T-PE of the MPLS-TP network.

Security reference model 1(a):

MPLS-TP network with Single Segment Pseudowire (SS-PW) from PE to PE. The trusted zone is PE1 to PE2 as illustrated in Figure 1.

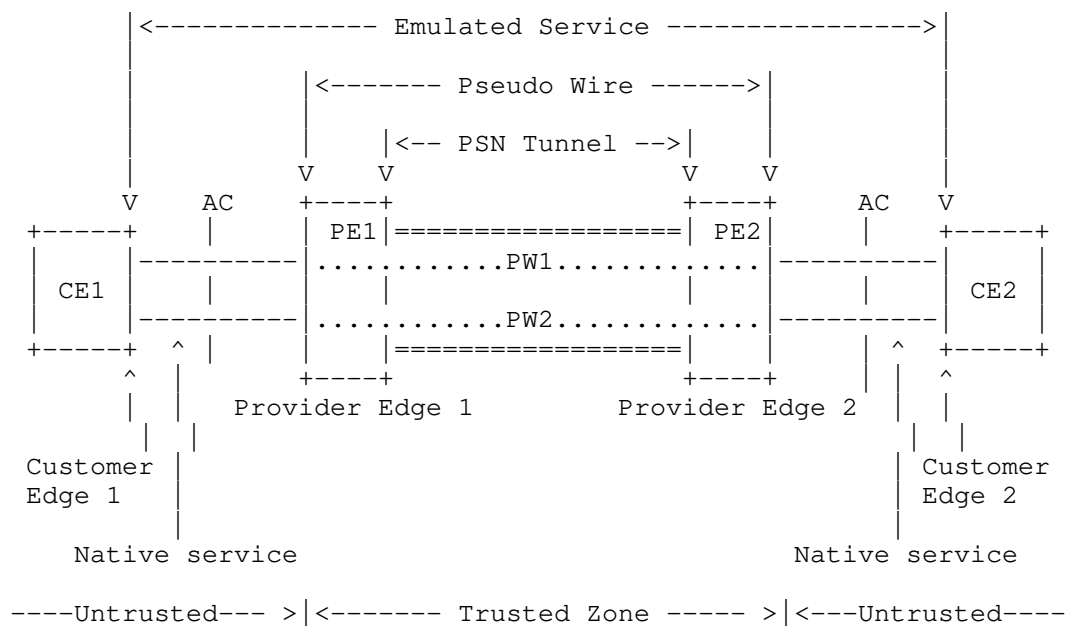


Figure 1: MPLS-TP Security Model 1 (a)

Security reference model 1(b) :

MPLS-TP network with Multi-Segment Pseudowire (MS-PW) from T-PE to T-PE. The trusted zone is T-PE1 to T-PE2 in this model as illustrated in Figure 2.

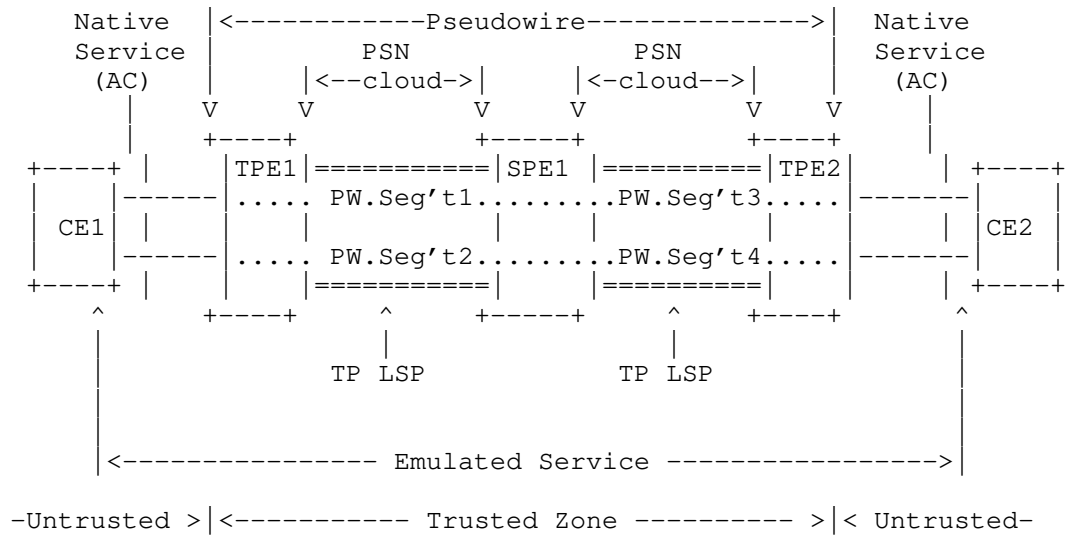


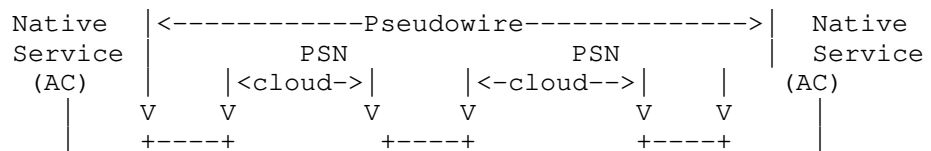
Figure 2: MPLS-TP Security Model 2 (b)

## 2.2. Security Reference Model 2

In the reference model 2, a single SP does not have the total control of PE/T-PE to PE/T-PE of the MPLS-TP network, S-PE and T-PE may be owned by different SPs or SPs and their customers. The MPLS-TP network is not contained in one trusted zone.

Security Reference Model 2(a)

MPLS-TP network with Multi-Segment Pseudowire (MS-PW) from PE to PE. The trusted zone is T-PE1 to S-PE, as illustrated in Figure 3.



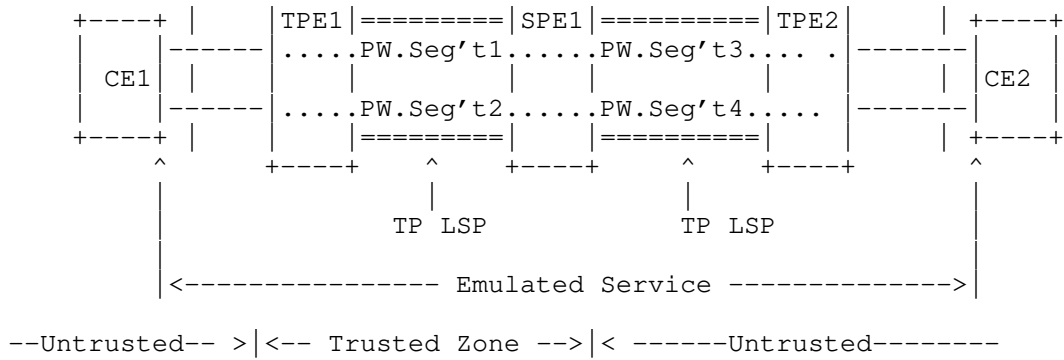


Figure 3: MPLS-TP Security Model 2(a)

#### Security Reference Model 2(b)

MPLS-TP network with Multi-Segment Pseudowire (MS-PW) from PE to PE. The trusted zone is S-PE, as illustrated in Figure 3.

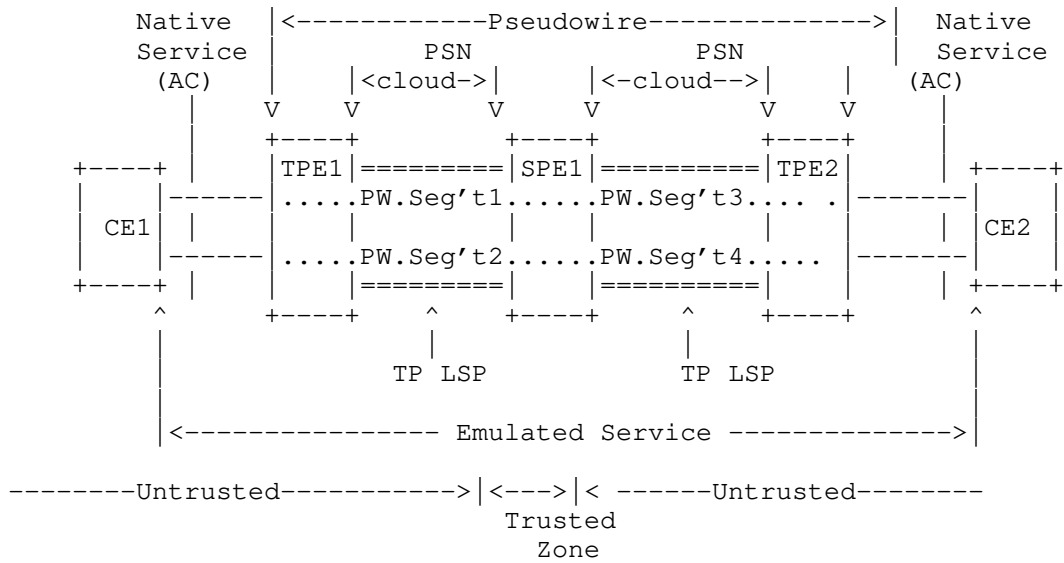


Figure 4: MPLS-TP Security Model 2(b)

Security Reference Model 2(c):

MPLS-TP network with Multi-Segment Pseudowire (MS-PW) from different Service Providers with PW inter-provider connections. The trusted zone is T-PE1 to S-PE3, as illustrated in Figure 5.

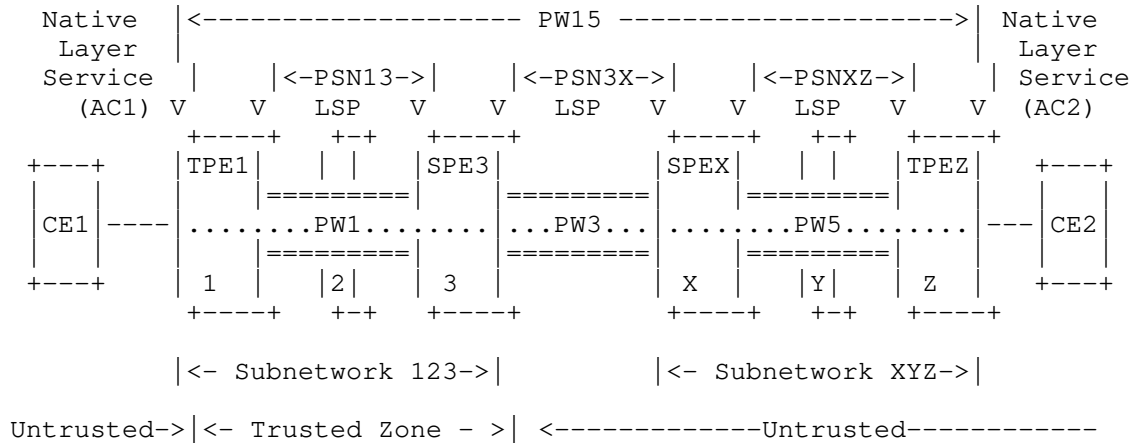
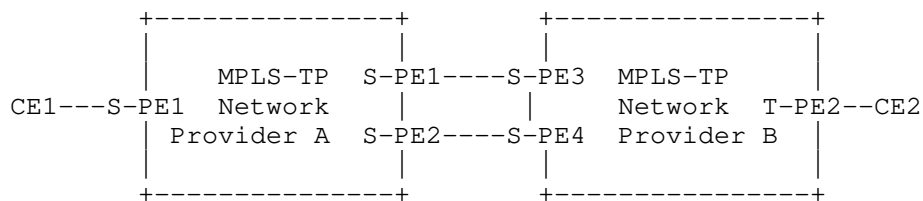


Figure 5: MPLS-TP Security Model 2(c)

The boundaries of a trust domain should be carefully defined when analyzing the security properties of each individual network, as illustrated from the above, the security boundaries determined which model would be applied to the use case analysis.

A key requirement of MPLS-TP networks is that the security of the trusted zone not be compromised by interconnecting the MPLS-TP, MPLS core infrastructure with another provider's core or T-PE devices, or end users.

In addition, neighbors may be trusted or untrusted. Neighbors may be authorized or unauthorized. Even though a neighbor may be authorized for communication, it may not be trusted. For example, when connecting with another provider's S-PE to set up Inter-AS LSPs, the other provider is considered an untrusted but may be authorized neighbor.



For Provider A:

Trusted Zone: Provider A MPLS-TP network

Trusted neighbors: T-PE1, S-PE1, S-PE2

Authorized but untrusted neighbor: provider B

Unauthorized neighbors: CE2

Figure 5. MPLS-TP trusted zone and authorized neighbor.

### 3. Security Requirements for MPLS-TP

This section covers security requirements for securing MPLS-TP network infrastructure. The MPLS-TP network can be operated without control plane or via dynamic control planes protocols. The security requirements related to new MPLS-TP OAM, recovery mechanisms, MPLS-TP and MPLS interconnection, and MPLS-TP specific operational requirements will be addressed in this section.

A service provider may choose the implementation options which are best fit for his/her network operation. This document does not state that a MPLS/GMPLS network must fulfill all security requirements listed to be secure.

These requirements are focused on: 1) how to protect the MPLS-TP network from various attacks originating outside the trusted zone including those from network users, both accidentally and maliciously; 2) prevention of operational errors resulted from misconfiguration within the trusted zone.

#### 3.1. Protection within the MPLS-TP Network

- MPLS-TP MUST support the physical and logical separation of data plane from the control plane and management plane. That is, if the control plane or/and management plane are attached and cannot function normally, the data plane should continue to forward packets without being impacted.
- MPLS-TP MUST support static provisioning of MPLS-TP LSP and PW with or without NMS/OSS, without using control protocols. This is particularly important in the case of security model 2(a) and 2(b) where the some or all T-PEs are not in the trusted zone, and in the inter-provider cases in security model 2(c) when the connecting S-PE is in the untrusted zone.
- MPLS-TP MUST support non-IP path options in addition to IP loopback option. Non-IP path option used in the model 2 may help to lower the potential risk of the S-PE/T-PE in the trusted zone to be attacked.
- MPLS-TP MUST support authentication of the any control protocol used for MPLS-TP network, as well as MPLS-TP network to dynamic MPLS network inter-connection.
- MPLS-TP MUST support mechanisms to prevent DOS attack through in-band OAM G-ACh/GAL.
- MPLS-TP MUST support hiding of the Service Provider infrastructure for all reference models regardless using static configuration or dynamic control plane.
- Security management requirements [MPLS-TP NM REQ]:
  - o MPLS-TP must support management communication channel security secure communication channels MUST be supported for all network traffic and protocols used to support management functions. This MUST include protocols used for configuration, monitoring, configuration backup, logging, time synchronization, authentication, and routing. The MCC MUST support application protocols that provide confidentiality and data integrity protection. Support the use of open cryptographic algorithms [RFC 3871]; Authentication - allow management connectivity and activity only from authenticated entities, and port access control.
  - o Distributed Denial of Service: It is possible to lessen the impact and potential for DoS and DDoS by using secure protocols, turning off unnecessary processes, logging and

monitoring, and ingress filtering. [RFC 4732] provides background on DOS in the context of the Internet.

(more to be added)

- Protection of Operational error

Due to the extensive use of static provisioning with or without NMS and OSS, the prevention of configuration errors should be addressed as major security requirements.

(to be added)

#### 4. Security Threats

This section discusses the various network security threats that may endanger MPLS-TP networks. The discussion is limited to those threats that are unique to MPLS-TP networks or that affect MPLS-TP network in unique ways.

A successful attack on a particular MPLS-TP network or on a SP's MPLS-TP infrastructure may cause one or more of the following ill effects:

1. Observation, modification, or deletion of a provider's or user's data, as well as replay or insertion of inauthentic data into a provider's or user's data stream, traffic pattern analysis.

These types of attacks apply to MPLS-TP traffic in a similar way of MPLS traffic regardless how the LSP is set up.

2. Attacks on GAL label, BFD messages:

- 1) GAL label or BFD label manipulation: including insertion of false label or messages, or modification, or removal the GAL labels or messages by attackers.

- 2) DOS attack through in-band OAM G-ACH/GAL, and BFD messages.

3. Disruption of a provider's and/or user's connectivity, or degradation of a provider's service quality.

- 1) Provider connectivity attacks:

- In the case of NMS is used for LSP set-up, the attacks would be through the attack of NMS.

- In the case of dynamic is used for dynamic provisioning, the attack would be on dynamic control plane. Most aspects are addressed in [RFC 5920].

- 2) User connectivity attack. This would be similar as PE/CE access attack in typical MPLS networks, addressed in [RFC 5920].

4. Probing a provider's network to determine its configuration, capacity, or usage.

These types of attack can happen through NMS attacks in the case of static provisioning, or through control plane attacks as in dynamic MPLS networks. It can also be combined attacks.

It is useful to consider that threats, whether malicious or accidental, may come from different categories of sources. For example they may come from:

- Other users whose services are provided by the same MPLS-TP core.
- The MPLS-TP SP or persons working for it.
- Other persons who obtain physical access to a MPLS-TP SP's site.
- Other persons who use social engineering methods to influence the behavior of a SP's personnel.
- Users of the MPLS-TP network itself.
- Others, e.g., attackers from the other sources, Internet if connected.
- Other SPs in the case of MPLS-TP Inter-provider connection. The provider may or may not be using MPLS-TP.
- Those who create, deliver, install, and maintain software for network equipment.

Given that security is generally a tradeoff between expense and risk, it is also useful to consider the likelihood of different attacks occurring. There is at least a perceived difference in the likelihood of most types of attacks being successfully mounted in different environments, such as:

- A MPLS-TP network inter-connecting with another provider's core
- A MPLS-TP configuration transiting the public Internet

Most types of attacks become easier to mount and hence more likely as the shared infrastructure via which service is provided expands from a single SP to multiple cooperating SPs to the global Internet. Attacks that may not be of sufficient likeliness to warrant concern in a closely controlled environment often merit defensive measures in broader, more open environments. In closed

communities, it is often practical to deal with misbehavior after the fact: an employee can be disciplined, for example.

The following sections discuss specific types of exploits that threaten MPLS-TP networks.

#### 4.1. Attacks on the Control Plane

- MPLS-TP LSP creation by an unauthorized element
- LSP message interception
- Attacks on G-Ach
- Attacks against LDP
- Attacks against RSVP-TE
- Attacks against GMPLS
- Denial of Service Attacks on the Network Infrastructure
- Attacks on the SP's MPLS/GMPLS Equipment via Management Interfaces
- Social Engineering Attacks on the SP's Infrastructure
- Cross-Connection of Traffic between Users
- Attacks against Routing Protocols
- Other Attacks on Control Traffic

#### 4.2. Attacks on the Data Plane

This category encompasses attacks on the provider's or end user's data. Note that from the MPLS-TP network end user's point of view, some of this might be control plane traffic, e.g. routing protocols running from user site A to user site B via IP or non-IP connections, which may be some type of VPN.

- Unauthorized Observation of Data Traffic
- Modification of Data Traffic
- Insertion of Inauthentic Data Traffic: Spoofing and Replay
- Unauthorized Deletion of Data Traffic

- Unauthorized Traffic Pattern Analysis
- Denial of Service Attacks
- Misconnection

## 5. Defensive Techniques for MPLS-TP Networks

The defensive techniques discussed in this document are intended to describe methods by which some security threats can be addressed. They are not intended as requirements for all MPLS-TP implementations. The MPLS-TP provider should determine the applicability of these techniques to the provider's specific service offerings, and the end user may wish to assess the value of these techniques to the user's service requirements. The operational environment determines the security requirements. Therefore, protocol designers need to provide a full set of security services, which can be used where appropriate.

The techniques discussed here include encryption, authentication, filtering, firewalls, access control, isolation, aggregation, and others.

### 5.1. Authentication

To prevent security issues arising from some DoS attacks or from malicious or accidental misconfiguration, it is critical that devices in the MPLS-TP should only accept connections or control messages from valid sources. Authentication refers to methods to ensure that message sources are properly identified by the MPLS-TP devices with which they communicate. This section focuses on identifying the scenarios in which sender authentication is required and recommends authentication mechanisms for these scenarios.

#### 5.1.1. Management System Authentication

Management system authentication includes the authentication of a PE to a centrally-managed network management or directory server when directory-based "auto-discovery" is used. It also includes authentication of a CE to the configuration server, when a configuration server system is used.

Authentication should be bi-directional, including PE or CE to configuration server authentication for PE or CE to be certain it is communicating with the right server.

#### 5.1.2. Peer-to-Peer Authentication

Peer-to-peer authentication includes peer authentication for network control protocols and other peer authentication (i.e., authentication of one IPsec security gateway by another).

Authentication should be bi-directional, including S-PE, T-PE, PE or CE to configuration server authentication for PE or CE to be certain it is communicating with the right server.

#### 5.1.3. Cryptographic Techniques for Authenticating Identity

Cryptographic techniques offer several mechanisms for authenticating the identity of devices or individuals. These include the use of shared secret keys, one-time keys generated by accessory devices or software, user-ID and password pairs, and a range of public-private key systems. Another approach is to use a hierarchical Certification Authority system to provide digital certificates.

#### 5.2. Access Control Techniques

##### - Access Control to Management Interfaces

Most of the security issues related to management interfaces can be addressed through the use of authentication techniques as described in the section on authentication. However, additional security may be provided by controlling access to management interfaces in other ways.

The Optical Internetworking Forum has done relevant work on protecting such interfaces with TLS, SSH, Kerberos, IPsec, WSS, etc. See OIF-SMI-01.0 "Security for Management Interfaces to Network Elements" [OIF-SMI-01.0], and "Addendum to the Security for Management Interfaces to Network Elements" [OIF-SMI-02.1]. See also the work in the ISMS WG.

Management interfaces, especially console ports on MPLS-TP devices, may be configured so they are only accessible out-of-band, through a system which is physically or logically separated from the rest of the MPLS-TP infrastructure.

Where management interfaces are accessible in-band within the MPLS-TP domain, filtering or firewalling techniques can be used to restrict unauthorized in-band traffic from having access to management interfaces. Depending on device capabilities, these

filtering or firewalling techniques can be configured either on other devices through which the traffic might pass, or on the individual MPLS-TP devices themselves.

### 5.3. Use of Isolated Infrastructure

One way to protect the infrastructure used for support of MPLS-TP is to separate the resources for support of MPLS-TP services from the resources used for other purposes

### 5.4. Use of Aggregated Infrastructure

In general, it is not feasible to use a completely separate set of resources for support of each service. In fact, one of the main reasons for MPLS-TP enabled services is to allow sharing of resources between multiple services and multiple users. Thus, even if certain services use a separate network from Internet services, nonetheless there will still be multiple MPLS-TP users sharing the same network resources.

In general, the use of aggregated infrastructure allows the service provider to benefit from stochastic multiplexing of multiple bursty flows, and also may in some cases thwart traffic pattern analysis by combining the data from multiple users. However, service providers must minimize security risks introduced from any individual service or individual users.

### 5.5. Service Provider Quality Control Processes

### 5.6. Verification of Connectivity

In order to protect against deliberate or accidental misconnection, mechanisms can be put in place to verify both end-to-end connectivity and hop-by-hop resources. These mechanisms can trace the routes of LSPs in both the control plane and the data plane.

## 6. Monitoring, Detection, and Reporting of Security Attacks

MPLS-TP network and service may be subject to attacks from a variety of security threats. Many threats are described in Section 3 of this document. Many of the defensive techniques described in this document and elsewhere provide significant levels of protection from a variety of threats. However, in addition to employing defensive techniques silently to protect against attacks, MPLS-TP services can also add value for both providers and

customers by implementing security monitoring systems to detect and report on any security attacks, regardless of whether the attacks are effective.

Attackers often begin by probing and analyzing defenses, so systems that can detect and properly report these early stages of attacks can provide significant benefits.

Information concerning attack incidents, especially if available quickly, can be useful in defending against further attacks. It can be used to help identify attackers or their specific targets at an early stage. This knowledge about attackers and targets can be used to strengthen defenses against specific attacks or attackers, or to improve the defenses for specific targets on an as-needed basis. Information collected on attacks may also be useful in identifying and developing defenses against novel attack types.

## 7. Security Considerations

Security considerations constitute the sole subject of this memo and hence are discussed throughout.

The document describes a variety of defensive techniques that may be used to counter the suspected threats. All of the techniques presented involve mature and widely implemented technologies that are practical to implement.

The document evaluates MPLS-TP security requirements from a customer's perspective as well as from a service provider's perspective. These sections re-evaluate the identified threats from the perspectives of the various stakeholders and are meant to assist equipment vendors and service providers, who must ultimately decide what threats to protect against in any given configuration or service offering.

## 8. IANA Considerations

This document contains no new IANA considerations.

## 9. Normative References

[RFC 5654], Niven-Jenkins, B., et al, "MPLS-TP Requirements", RFC 5654, September 2009.

MPLS-TP Security framework  
October 2010

[RFC 3871] Jones, G., "Operational Security Requirements for Large Internet Service Provider (ISP) IP Network Infrastructure", RFC 3871, September 2004.

[RFC 4732] Handley, M., et al, "Internet Denial-of-Service Considerations", RFC 4732, November 2006.

#### 10. Informative References

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

[OIF-SMI-01.0] Renee Esposito, "Security for Management Interfaces to Network Elements", Optical Internetworking Forum, Sept. 2003.

[OIF-SMI-02.1] Renee Esposito, "Addendum to the Security for Management Interfaces to Network Elements", Optical Internetworking Forum, March 2006.

[RFC3631] S. Bellovin, C. Kaufman, J. Schiller, "Security Mechanisms for the Internet," December 2003.

[MFA MPLS ICI] N. Bitar, "MPLS InterCarrier Interconnect Technical Specification", IP/MPLS Forum 19.0.0, April 2008.

[RFC 5921] Bocci, M., Bryant, et al., "A Framework for MPLS in Transport Networks", July 2010.

[opsec efforts] C. Lonvick and D. Spak, "Security Best Practices Efforts and Documents", draft-ietf-opsec-efforts-08.txt, June 2008.

[RFC 5920] L. Fang, et al, Security Framework for MPLS and GMPLS Networks, July 2010.

[MPLS-TP NM REQ] Hing-Kam Lam, Scott Mansfield, Eric Gray, MPLS TP Network Management Requirements, draft-ietf-mpls-tp-nm-req-06.txt, October 2009.

#### 11. Author's Addresses

Luyuan Fang  
Cisco Systems, Inc.  
300 Beaver Brook Road  
Boxborough, MA 01719  
USA

MPLS-TP Security framework  
October 2010

Email: lufang@cisco.com

Ben Niven-Jenkins  
Velocix  
326 Cambridge Science Park  
Milton Road,  
Cambridge  
CB4 0WG, UK

Email: ben@niven-jenkins.co.uk

Raymond Zhang  
British Telecom  
BT Center  
81 Newgate Street  
London, EC1A 7AJ  
United Kingdom  
Email: raymond.zhang@bt.com

Nabil Bitar  
Verizon  
40 Sylvan Road  
Waltham, MA 02145  
USA  
Email: nabil.bitar@verizon.com

Masahiro DAIKOKU  
KDDI corporation  
3-11-11.Iidabashi, Chiyodaku, Tokyo  
Japan  
Email: ms-daikoku@kddi.com

SCOTT MANSFIELD  
Ericsson  
Email: scott.mansfield@ericsson.com

Lai Wang  
Telenor  
Telenor Norway  
Office Snaroyveien  
1331 Fornebu  
Email: Lai.wang@telenor.com

Network Working Group  
Internet Draft  
Intended status: Informational  
Expires: April 25, 2011

Luyuan Fang  
Dan Frost  
Cisco Systems  
Nabil Bitar  
Verizon  
Raymond Zhang  
BT  
Masahiro DAIKOKU  
KDDI  
Jian Ping Zhang  
China Telecom, Shanghai  
Lei Wang  
Telenor  
Mach(Guoyi) Chen  
Huawei Technologies  
Nurit Sprecher  
Nokia Siemens Networks

October 25, 2010

MPLS-TP Use Cases Studies and Design Considerations  
draft-fang-mpls-tp-use-cases-and-design-02.txt

#### Abstract

This document provides use case studies and network design considerations for Multiprotocol Label Switching Transport Profile (MPLS-TP).

In the recent years, MPLS-TP has emerged as the technology of choice to meet the needs of transport evolution. Many service providers (SPs) intend to replace SONET/SDH, TDM, ATM traditional transport technologies with MPLS-TP, to achieve higher efficiency, lower operational cost, while maintaining transport characteristics. The use cases for MPLS-TP include Mobile backhaul, Metro Ethernet access and aggregation, and packet optical transport. The design considerations include operational experience, standards compliance, technology maturity, end-to-end forwarding and OAM consistency, compatibility with IP/MPLS networks, and multi-vendor interoperability. The goal is to provide reliable, manageable, and scalable transport solutions.

The unified MPLS strategy, using MPLS from core to aggregation and access (e.g. IP/MPLS in the core, IP/MPLS or MPLS-TP in aggregation and access) appear to be very attractive to many SPs. It streamlines the operation, many help to reduce the overall complexity and

improve end-to-end convergence. It leverages the MPLS experience, and enhances the ability to support revenue generating services.

#### 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 January 12, 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 in effect on the date of publication of this document (<http://trustee.ietf.org/license-info>). Please review these documents carefully, as they describe your rights and restrictions with respect to this document.

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.....	4
1.1. Background and Motivation.....	4
1.2. Contributing authors.....	5
2. Terminologies.....	5
3. Overview of MPLS-TP base functions.....	6
3.1. MPLS-TP development principles.....	6
3.2. Data Plane.....	7
3.3. Control Plane.....	7
3.4. OAM.....	7
3.5. Survivability.....	8
4. MPLS-TP Use Case Studies.....	8
4.1. Mobile Backhaul.....	8
4.2. Metro Access and Aggregation.....	10
4.3. Packet Optical Transport.....	10
5. Network Design Considerations.....	11
5.1. IP/MPLS vs. MPLS-TP.....	11
5.2. Standards compliance.....	11
5.3. End-to-end MPLS OAM consistency.....	12
5.4. Delay and delay variation.....	12
5.5. General network design considerations.....	15
6. MPLS-TP Deployment Consideration.....	15
6.1. Network Modes Selection.....	15
6.2. Provisioning Modes Selection.....	16
7. Security Considerations.....	16
8. IANA Considerations.....	16
9. Normative References.....	17
10. Informative References.....	17
11. Author's Addresses.....	17

## Requirements Language

Although this document is not a protocol specification, 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 [RFC 2119].

## 1. Introduction

### 1.1. Background and Motivation

This document provides case studies and network design considerations for Multiprotocol Label Switching Transport Profile (MPLS-TP).

In recent years, the urgency for moving from traditional transport technologies such as SONET/SDH, TDM/ATM to new packet technologies has been rising. This is largely due to the tremendous success of data services, such as IPTV and IP Video for content downloading, streaming, and sharing; rapid growth of mobile services, especially smart phone applications; business VPNs and residential broadband. Continued network convergence effort is another contributing factor for transport moving toward packet technologies. After several years of heated debate, MPLS-TP has emerged as the next generation transport technology of choice for many service providers worldwide.

MPLS-TP is based on MPLS technologies. MPLS-TP re-use a subset of MPLS base functions, such as MPLS data forwarding, Pseudo-wire encapsulation for circuit emulation, and GMPLS for control plane option; MPLS-TP extended current MPLS OAM functions, such as BFD extension for Connectivity for proactive Connectivity Check (CC) and Connectivity Verification (CV), and Remote Defect Indication (RDI), LSP Ping Extension for on demand Connectivity Check (CC) and Connectivity Verification (CV), fault allocation, and remote integrity check. New tools are being defined for alarm suppression with Alarm Indication Signal (AIS), and trigger of switch over with Link Defect Indication (LDI). The goal is to take advantage of the maturity of MPLS technology, re-use the existing component when possible and extend the existing protocols or create new procedures/protocols when needed to fully satisfy the transport requirements.

The general requirements of MPLS-TP are provided in MPLS-TP Requirements [RFC 5654], and the architectural framework are defined in MPLS-TP Framework [RFC 5921]. This document intent to provide the use case studies and design considerations from practical point of view based on Service Providers deployments plans and field implementations.

The most common use cases for MPLS-TP include Mobile Backhaul, Metro Ethernet access and aggregation, and Packet Optical Transport. MPLS-TP data plane architecture, path protection mechanisms, and OAM functionalities are used to support these deployment scenarios.

As part of MPLS family, MPLS-TP complements today's IP/MPLS technologies; it closes the gaps in the traditional access and aggregation transport to provide end-to-end solutions in a cost efficient, reliable, and interoperable manner.

The unified MPLS strategy, using MPLS from core to aggregation and access (e.g. IP/MPLS in the core, IP/MPLS or MPLS-TP in aggregation and access) appear to be very attractive to many SPs. It streamlines the operation, many help to reduce the overall complexity and improve end-to-end convergence. It leverages the MPLS experience, and enhances the ability to support revenue generating services.

The design considerations discussed in this document are generic. While many design criteria are commonly apply to most of SPs, each individual SP may place the importance of one aspect over another depending on the existing operational environment, the applications need to be supported, the design objective, and the expected duration of the network to be in service for a particular design.

## 1.2. Contributing authors

Luyuan Fang, Cisco Systems  
Nabil Bitar, Verizon  
Raymond Zhang, BT  
Masahiro DAIKOKU, KDDI  
Jian Ping Zhang, China Telecom, Shanghai  
Mach(Guoyi) Chen, Huawei Technologies

## 2. Terminologies

AIS	Alarm Indication Signal
APS	Automatic Protection Switching
ATM	Asynchronous Transfer Mode
BFD	Bidirectional Forwarding Detection
CC	Continuity Check
CE	Customer Edge device
CV	Connectivity Verification
CM	Configuration Management
DM	Packet delay measurement
ECMP	Equal Cost Multi-path
FM	Fault Management
GAL	Generic Alert Label
G-ACH	Generic Associated Channel
GMPLS	Generalized Multi-Protocol Label Switching
LB	Loopback

LDP	Label Distribution Protocol
LM	Packet loss measurement
LSP	Label Switched Path
LT	Link trace
MEP	Maintenance End Point
MIP	Maintenance Intermediate Point
MP2MP	Multi-Point to Multi-Point connections
MPLS	Multi-Protocol Label Switching
MPLS-TP	MPLS transport profile
OAM	Operations, Administration, and Management
P2P	Point to Multi-Point connections
P2MP	Point to Point connections
PE	Provider-Edge device
PHP	Penultimate Hop Popping
PM	Performance Management
PW	Pseudowire
RDI	Remote Defect Indication
RSVP-TE	Resource Reservation Protocol with Traffic Engineering
Extensions	
SLA	Service Level Agreement
SNMP	Simple Network Management Protocol
SONET	Synchronous Optical Network
S-PE	Switching Provider Edge
SRLG	Shared Risk Link Group
TDM	Time Division Multiplexing
TE	Traffic Engineering
TTL	Time-To-Live
T-PE	Terminating Provider Edge
VPN	Virtual Private Network

### 3. Overview of MPLS-TP base functions

The section provides a summary view of MPLS-TP technology, especially in comparison to the base IP/MPLS technologies. For complete requirements and architecture definitions, please refer to [RFC 5654] and [RFC 5921].

#### 3.1. MPLS-TP development principles

The principles for MPLS-TP development are: meeting transport requirements; maintain transport characteristics; re-using the existing MPLS technologies wherever possible to avoid duplicate the effort; ensuring consistency and inter-operability of MPLS-TP and IP/MPLS networks; developing new tools as necessary to fully meet transport requirements.

MPLS-TP Technologies include four major areas: Data Plane, Control Plane, OAM, and Survivability. The short summary is provided below.

### 3.2. Data Plane

MPLS-TP re-used MPLS and PW architecture; and MPLS forwarding mechanism;

MPLS-TP extended the LSP support from unidirectional to both bi-directional unidirectional support.

MPLS-TP defined PHP as optional, disallowed ECMP and MP2MP, only P2P and P2MP are allowed.

### 3.3. Control Plane

MPLS-TP allowed two control plane options:

Static: Using NMS for static provisioning;

Dynamic Control Plane using GMPLS, OSPF-TE, RSVP-TE for full automation

ACH concept in PW is extended to GACH for MPLS-TP LSP to support in-band OAM.

Both Static and dynamic control plane options must allow control plane and data plane separation.

### 3.4. OAM

OAM received most attention in MPLS-TP development; Many OAM functions require protocol extensions or new development to meet the transport requirements.

1) Continuity Check (CC), Continuity Verification (CV), and Remote Integrity:

- Proactive CC and CV: Extended BFD
- On demand CC and CV: Extended LSP Ping
- Proactive Remote Integrity: Extended BFD
- On demand Remote Integrity: Extended LSP Ping

2) Fault Management:

- Fault Localization: Extended LSP Ping
- Alarm Suppression: create AIS
- Remote Defect Indication (RDI): Extended BFD
- Lock reporting: Create Lock Instruct
- Link defect Indication: Create LDI

- Static PW defect indication: Use Static PW status

Performance Management:

- Loss Management: Create MPLS-TP loss/delay measurement
- Delay Measurement: Create MPLS-TP loss/delay measurement

### 3.5. Survivability

- Deterministic path protection
- Switch over within 50ms
- 1:1, 1+1, 1:N protection
- Linear protection
- Ring protection

## 4. MPLS-TP Use Case Studies

### 4.1. Mobile Backhaul

Mobility is one of the fastest growing areas in communication world wide. For some regions, the tremendous rapid mobile growth is fueled with lack of existing land-line and cable infrastructure. For other regions, the introduction of Smart phones quickly drove mobile data traffic to become the primary mobile bandwidth consumer, some SPs have already seen 85% of total mobile traffic are data traffic.

MPLS-TP has been viewed as a suitable technology for Mobile backhaul.

#### 4.1.1. 2G and 3G Mobile Backhaul Support

MPLS-TP is commonly viewed as a very good fit for 2G)/3G Mobile backhaul.

2G (GSM/CDMA) and 3G (UMTS/HSPA/1xEVDO) Mobile Backhaul Networks are dominating mobile infrastructure today.

The connectivity for 2G/3G networks are Point to point. The logical connections are hub-and-spoke. The physical construction of the networks can be star topology or ring topology. In the Radio Access Network (RAN), each mobile base station (BTS/Node B) is communicating with one Radio Controller (BSC/RNC) only. These connections are often statically set up.

Hierarchical Aggregation Architecture / Centralized Architecture are often used for pre-aggregation and aggregation layers. Each aggregation networks inter-connects with multiple access networks.

For example, single aggregation ring could aggregate traffic for 10 access rings with total 100 base stations.

The technology used today is largely ATM based. Mobile providers are replacing the ATM RAN infrastructure with newer packet technologies. IP RAN networks with IP/MPLS technologies are deployed today by many SPs with great success. MPLS-TP is another suitable choice for Mobile RAN. The P2P connection from base station to Radio Controller can be set statically to mimic the operation today in many RAN environments, in-band OAM and deterministic path protection would support the fast failure detection and switch over to satisfy the SLA agreement. Bidirectional LSP may help to simplify the provisioning process. The deterministic nature of MPLS-TP LSP set up can also help packet based synchronization to maintain predictable performance regarding packet delay and jitters.

#### 4.1.2. LTE Mobile Backhaul

One key difference between LTE and 2G/3G Mobile networks is that the logical connection in LTE is mesh while 2G/3G is P2P star connections.

In LTE, the base stations eNB/BTS can communicate with multiple Network controllers (PSW/SGW or ASNGW), and each Radio element can communicate with each other for signal exchange and traffic offload to wireless or Wireline infrastructures.

IP/MPLS may have a great advantage in any-to-any connectivity environment. The use of mature IP or L3VPN technologies is particularly common in the design of SP's LTE deployment plan.

MPLS-TP can also bring advantages with the in-band OAM and path protection mechanism. MPLS-TP dynamic control-plane with GMPLS signaling may bring additional advantages in the mesh environment for real time adaptivities, dynamic topology changes, and network optimization.

Since MPLS-TP is part of the MPLS family. Many component already shared by both IP/MPLS and MPLS-TP, the line can be further blurred by sharing more common features. For example, it is desirable for many SPs to introduce the in-band OAM developed for MPLS-TP back into IP/MPLS networks as an enhanced OAM option. Today's MPLS PW can also be set statically to be deterministic if preferred by the SPs without going through full MPLS-TP deployment.

#### 4.1.3. WiMAX Backhaul

WiMAX Mobile backhaul shares the similar characteristics as LTE, with mesh connections rather than P2P, star logical connections.

#### 4.2. Metro Access and Aggregation

Some SPs are building new Access and aggregation infrastructure, while others plan to upgrade/replace of existing transport infrastructure with new packet technologies such as MPLS-TP. The later is of course more common than the former.

The access and aggregation networks today can be based on ATM, TDM, MSTP, or Ethernet technologies as later development.

Some SPs announced their plans for replacing their ATM or TDM aggregation networks with MPLS-TP technologies, because the ATM / TDM aggregation networks are no longer suited to support the rapid bandwidth growth, and they are expensive to maintain or may also be and impossible expand due to End of Sale and End of Life legacy equipments. The statistical muxing in MPLS-TP helps to achieve higher efficiency comparing with the time division scheme in the legacy technologies.

The unified MPLS strategy, using MPLS from core to aggregation and access (e.g. IP/MPLS in the core, IP/MPLS or MPLS-TP in aggregation and access) appear to be very attractive to many SPs. It streamlines the operation, many help to reduce the overall complexity and improve end-to-end convergence. It leverages the MPLS experience, and enhances the ability to support revenue generating services.

The current requirements from the SPs for ATM/TDM aggregation replacement often include maintaining the current operational model, with the similar user experience in NMS, supports current access network (e.g. Ethernet, ADSL, ATM, STM, etc.), support the connections with the core networks, support the same operational feasibility even after migrating to MPLS-TP from ATM/TDM and services (OCN, IP-VPN, E-VLAN, Dedicated line, etc.). MPLS-TP currently defined in IETF are meeting these requirements to support a smooth transition.

The green field network deployment is targeting using the state of art technology to build most stable, scalable, high quality, high efficiency networks to last for the next many years. IP/MPLS and MPLS-TP are both good choices, depending on the operational model.

#### 4.3. Packet Optical Transport

(to be added)

## 5. Network Design Considerations

### 5.1. IP/MPLS vs. MPLS-TP

Questions we often hear: I have just built a new IP/MPLS network to support multi-services, including L2/L3 VPNs, Internet service, IPTV, etc. Now there is new MPLS-TP development in IETF. Do I need to move onto MPLS-TP technology to state current with technologies?

The answer is no generally speaking. MPLS-TP is developed to meet the needs of traditional transport moving towards packet. It is geared to support the transport behavior coming with the long history. IP/MPLS and MPLS-TP both are state of art technologies. IP/MPLS support both transport (e.g. PW, RSVP-TE, etc.) and services (e.g. L2/L3 VPNs, IPTV, Mobile RAN, etc.), MPLS-TP provides transport only. The new enhanced OAM features built in MPLS-TP should be share in both flavors through future implementation.

Another question: I need to evolve my ATM/TDM/SONET/SDH networks into new packet technologies, but my operational force is largely legacy transport, not familiar with new data technologies, and I want to maintain the same operational model for the time being, what should I do? The answer would be: MPLS-TP may be the best choice today for the transition.

A few important factors need to be considered for IP/MPLS or MPLS-TP include:

- Technology maturity (IP/MPLS is much more mature with 12 years development)
- Operation experience (Work force experience, Union agreement, how easy to transition to a new technology? how much does it cost?)
- Needs for Multi-service support on the same node (MPLS-TP provide transport only, does not replace many functions of IP/MPLS)
- LTE, IPTV/Video distribution considerations (which path is the most viable for reaching the end goal with minimal cost? but it also meet the need of today's support)

### 5.2. Standards compliance

It is generally recognized by SPs that standards compliance are important for driving the cost down and product maturity up, multi-vendor interoperability, also important to meet the expectation of the business customers of SP's.

MPLS-TP is a joint work between IETF and ITU-T. In April 2008, IETF and ITU-T jointly agreed to terminate T-MPLS and progress MPLS-TP as

joint work [RFC 5317]. The transport requirements would be provided by ITU-T, the protocols would be developed in IETF.

T-MPLS is not MPLS-TP. T-MPLS solution would not inter-op with IP/MPLS, it would not be compatible with MPLS-TP defined in IETF.

### 5.3. End-to-end MPLS OAM consistency

In the case Service Providers deploy end-to-end MPLS solution with the combination of dynamic IP/MPLS and static or dynamic MPLS-TP cross core, service edge, and aggregation/access networks, end-to-end MPLS OAM consistency becomes an essential requirements from many Service Provider. The end-to-end MPLS OAM can only be achieved through implementation of IETF MPLS-TP OAM definitions.

### 5.4. Delay and delay variation

Background/motivation: Telecommunication Carriers plan to replace the aging TDM Services (e.g. legacy VPN services) provided by Legacy TDM technologies/equipments to new VPN services provided by MPLS-TP technologies/equipments with minimal cost. The Carriers cannot allow any degradation of service quality, service operation Level, and service availability when migrating out of Legacy TDM technologies/equipments to MPLS-TP transport. The requirements from the customers of these carriers are the same before and after the migration.

#### 5.4.1. Network Delay

From our recent observation, more and more Ethernet VPN customers becoming very sensitive to the network delay issues, especially the financial customers. Many of those customers has upgraded their systems in their Data Centers, e.g., their accounting systems. Some of the customers built the special tuned up networks, i.e. Fiber channel networks, in their Data Centers, this tripped more strict delay requirements to the carriers.

There are three types of network delay:

##### 1. Absolute Delay Time

Absolute Delay Time here is the network delay within SLA contract. It means the customers have already accepted the value of the Absolute Delay Time as part of the contract before the Private Line Service is provisioned.

## 2. Variation of Absolute Delay Time (without network configuration changes).

The variation under discussion here is mainly induced by the buffering in network elements.

Although there is no description of Variation of Absolute Delay Time on the contract, this has no practical impact on the customers who contract for the highest quality of services available. The bandwidth is guaranteed for those customers' traffic.

## 3. Relative Delay Time

Relative Delay Time is the difference of the Absolute Delay Time between using working and protect path.

Ideally, Carriers would prefer the Relative Delay Time to be zero, for the following technical reasons and network operation feasibility concerns.

The following are the three technical reasons:

### Legacy throughput issue

In the case that Relative Delay Time is increased between FC networks or TCP networks, the effective throughput is degraded. The effective throughput, though it may be recovered after revert back to the original working path in revertive mode.

On the other hand, in that case that Relative Delay Time is decreased between FC networks or TCP networks, buffering over flow may occur at receiving end due to receiving large number of busty packets. As a consequence, effective throughput is degraded as well. Moreover, if packet reordering is occurred due to RTT decrease, unnecessary packet resending is induced and effective throughput is also further degraded. Therefore, management of Relative Delay Time is preferred, although this is known as the legacy TCP throughput issue.

### Locating Network Accelerators at CE

In order to improve effective throughput between customer's FC networks over Ethernet private line service, some customer put "WAN Accelerator" to increase throughput value. For example, some WAN Accelerators at receiving side may automatically send back "R\_RDY" in order to avoid decreasing a number of BBcredit at sending side, and the other WAN Accelerators at sending side may have huge number of initial BB credit.

When customer tunes up their CE by locating WAN Accelerator, for example, when Relative Delay Time is changes, there is a possibility that effective throughput is degraded. This is because a lot of packet destruction may be occurred due to loss of synchronization, when change of Relative delay time induces packet reordering. And, it is difficult to re-tune up their CE network element automatically when Relative Delay Time is changed, because only less than 50 ms network down detected at CE.

Depending on the tuning up method, since Relative Delay Time affects effective throughput between customer's FC networks, management of Relative Delay Time is preferred.

c) Use of synchronized replication system

Some strict customers, e.g. financial customers, implement "synchronized replication system" for all data back-up and load sharing. Due to synchronized replication system, next data processing is conducted only after finishing the data saving to both primary and replication DC storage. And some tuning function could be applied at Server Network to increase throughput to the replication DC and Client Network. Since Relative Delay Time affects effective throughput, management of Relative Delay Time is preferred.

The following are the network operational feasibility issues.

Some strict customers, e.g., financial customer, continuously checked the private line connectivity and absolute delay time at CEs. When the absolute delay time is changed, that is Relative delay time is increased or decreased, the customer would complain.

From network operational point of view, carrier want to minimize the number of customers complains, MPLS-TP LSP provisioning with zero Relative delay time is preferred and management of Relative Delay Time is preferred.

Obviously, when the Relative Delay Time is increased, the customer would complain about the longer delay. When the Relative Delay Time is decreased, the customer expects to keep the lesser Absolute Delay Time condition and would complain why Carrier did not provide the best solution in the first place. Therefore, MPLS-TP LSP provisioning with zero Relative Delay Time is preferred and management of Relative Delay Time is preferred.

More discussion will be added on how to manage the Relative delay time.

### 5.5. General network design considerations

- Migration considerations
- Resiliency
- Scalability
- Performance

## 6. MPLS-TP Deployment Consideration

### 6.1. Network Modes Selection

When considering deployment of MPLS-TP in the network, possibly couple of questions will come into mind, for example, where should the MPLS-TP be deployed? (e.g., access, aggregation or core network?) Should IP/MPLS be deployed with MPLS-TP simultaneously? If MPLS-TP and IP/MPLS is deployed in the same network, what is the relationship between MPLS-TP and IP/MPLS (e.g., peer or overlay?) and where is the demarcation between MPLS-TP domain and IP/MPLS domain? The results for these questions depend on the real requirements on how MPLS-TP and IP/MPLS are used to provide services. For different services, there could be different choice. According to the combination of MPLS-TP and IP/MPLS, here are some typical network modes:

Pure MPLS-TP as the transport connectivity (E2E MPLS-TP), this situation more happens when the network is a totally new constructed network. For example, a new constructed packet transport network for Mobile Backhaul, or migration from ATM/TDM transport network to packet based transport network.

Pure IP/MPLS as transport connectivity (E2E IP/MPLS), this is the current practice for many deployed networks.

MPLS-TP combines with IP/MPLS as the transport connectivity (Hybrid mode)

Peer mode, some domains adopt MPLS-TP as the transport connectivity; other domains adopt IP/MPLS as the transport connectivity. MPLS-TP domains and IP/MPLS domains are interconnected to provide transport connectivity. Considering there are a lot of IP/MPLS deployments in the field, this mode may be the normal practice in the early stage of MPLS-TP deployment.

Overlay mode

b-1: MPLS-TP as client of IP/MPLS, this is for the case where MPLS-TP domains are distributed and IP/MPLS do-main/network is used for the connection of the distributed MPLS-TP domains. For examples, there are some service providers who have no their own Backhaul network, they have to rent the Backhaul network that is IP/MPLS based from other service providers.

b-2: IP/MPLS as client of MPLS-TP, this is for the case where transport network below the IP/MPLS network is a MPLS-TP based network, the MPLS-TP network provides transport connectivity for the IP/MPLS routers, the usage is analogous as today's ATM/TDM/SDH based transport network that are used for providing connectivity for IP/MPLS routers.

#### 6.2. Provisioning Modes Selection

As stated in MPLS-TP requirements [RFC5654], MPLS-TP network MUST be possible to work without using Control Plane. And this does not mean that MPLS-TP network has no control plane. Instead, operators could deploy their MPLS-TP with static provisioning (e.g., CLI, NMS etc.), dynamic control plane signaling (e.g., OSPF-TE/ISIS-TE, GMPLS, LDP, RSVP-TE etc.), or combination of static and dynamic provisioning (Hybrid mode). Each mode has its own pros and cons and how to determine the right mode for a specific network mainly depends on the operators' preference. For the operators who are used to operate traditional transport network and familiar with the Transport-Centric operational model (e.g., NMS configuration without control plane) may prefer static provisioning mode. The dynamic provisioning mode is more suitable for the operators who are familiar with the operation and maintenance of IP/MPLS network where a fully dynamic control plane is used. The hybrid mode may be used when parts of the network are provisioned with static way and the other parts are controlled by dynamic signaling. For example, for big SP, the network is operated and maintained by several different departments who prefer to different modes, thus they could adopt this hybrid mode to support both static and dynamic modes hence to satisfy different requirements. Another example is that static provisioning mode is suitable for some parts of the network and dynamic provisioning mode is suitable for other parts of the networks (e.g., static for access network, dynamic for metro aggregate and core network).

Note: This draft is work in progress, more would be filled in the following revision.

#### 7. Security Considerations

Reference to [RFC 5920]. More will be added.

#### 8. IANA Considerations

This document contains no new IANA considerations.

## 9. Normative References

[RFC 5317]: Joint Working Team (JWT) Report on MPLS Architectural Considerations for a Transport Profile, Feb. 2009.

[RFC 5654], Niven-Jenkins, B., et al, "MPLS-TP Requirements," RFC 5654, September 2009.

(More to be added)

## 10. Informative References

[RFC 5921] Bocci, M., ED., Bryant, S., ED., et al., Frost, D. ED., Levrau, L., Berger., L., "A Framework for MPLS in Transport," July 2010.

[RFC 5920] L. Fang, ED., et al, "Security Framework for MPLS and GMPLS Networks, " July 2010.

(More to be added)

## 11. Author's Addresses

Luyuan Fang  
Cisco Systems, Inc.  
111 Wood Ave. South  
Iselin, NJ 08830  
USA  
Email: lufang@cisco.com

Dan Frost  
Cisco Systems, Inc.  
Email: danfrost@cisco.com

Nabil Bitar  
Verizon  
40 Sylvan Road  
Waltham, MA 02145  
USA  
Email: nabil.bitar@verizon.com

Raymond Zhang  
British Telecom

BT Center  
81 Newgate Street  
London, EC1A 7AJ  
United Kingdom  
Email: raymond.zhang@bt.com

Masahiro DAIKOKU  
KDDI corporation  
3-11-11.Iidabashi, Chiyodaku, Tokyo  
Japan  
Email: ms-daikoku@kddi.com

Jian Ping Zhang  
China Telecom, Shanghai  
Room 3402, 211 Shi Ji Da Dao  
Pu Dong District, Shanghai  
China  
Email: zhangjp@shtel.com.cn

Lai Wang  
Telenor  
Telenor Norway  
Office Snaroyveien  
1331 Fornebu  
Email: Lai.wang@telenor.com

Mach(Guoyi) Chen  
Huawei Technologies Co., Ltd.  
No. 3 Xixi Road  
Shangdi Information Industry Base  
Hai-Dian District, Beijing 100085  
China  
Email: mach@huawei.com

Nurit Sprecher  
Nokia Siemens Networks  
3 Hanagar St. Neve Ne'eman B  
Hod Hasharon, 45241  
Israel  
Email: nurit.sprecher@nsn.com

MPLS Working Group  
Internet Draft  
Intended status: Standards Track  
Expires: November 2010

Feng Huang (Editor)  
Lieven Levrau (Editor)  
Alcatel-Lucent

Han LI (Editor)  
China Mobile

Ruiquan Jing (Editor)  
China Telecom

May 11, 2010

Diagnostic tool-test for MPLS transport profile  
draft-flh-mpls-tp-oam-diagnostic-test-01

Abstract

This document describes a Multi-Protocol Label Switching Transport Profile (MPLS-TP) Operations, Administration and Maintenance (OAM) diagnostic tool-TST (test), which is used to perform one-way, or two way on-demand out-of-service measuring throughput or in-service diagnostics tests for verifying throughput.

Conventions used in 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 RFC 2119 [1].

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 November 11, 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 BSD License.

#### Table of Contents

1. Introduction.....	3
1.1. Authors.....	3
2. Terminology.....	3
3. Mechanics of TST.....	4
3.1. General Requirements.....	4
3.2. Transmission.....	5
3.3. Receive.....	7
3.4. Performance Monitoring counter and throughput calculation	7
4. TST Frame format (PDU).....	7
5. Security Considerations.....	10
6. IANA Considerations.....	10
7. Acknowledgments.....	10
8. References.....	11
8.1. Normative References.....	11
8.2. Informative References.....	11
Authors' Addresses.....	11

## 1. Introduction

MPLS-TP is technology of packet transport network, which requirement is defined in MPLS-TP requirement [2], and OAM is its most important function. MPLS-TP OAM requirement [3] define diagnostic tools that MAY be used for PW, LSP and Section, such as consists in looping the traffic at an Intermediate Point or End point back to the originating End Point-loopback. And another example of such diagnostic tool-test (TST) consists in estimating the bandwidth or throughput of transport path e.g., an LSP.

This document defines one diagnostic tool-TST (test) for bandwidth estimating, measuring or verifying. And it describes TST OAM frame format and the procedures for the transmission, receive of such OAM frames.

The TST function SHOULD be performed between End Points of PWs, LSPs and Sections.

### 1.1. Authors

Feng Huang, Lieven Levrau, Han Li and Ruiquan Jing.

## 2. Terminology

CRC	Cyclic Redundancy Check
G-ACh	Generic Associated Channel
ACH	Associated Channel Header
ITU-T	International telecom union-Telecom
LCK	lock
LSP	Label Switch Path
MEP	ME Edge Point
MIP	ME Intermediated Point
MPLS-TP	MPLS transport profile
OAM	Operations Administration and Maintenance

PDU      Payload Data Unit

PRBS     pseudo-random code stream

PW       Pseudo wire

TLV      Type Length Value

TST      Test

### 3. Mechanics of TST

#### 3.1. General Requirements

The proposed test tool can be used to perform one-way or two way on-demand in-service or out-of-service diagnostics tests, which include verifying throughput. Throughput in this document refers a capacity in terms of line rate; it is the amount of bits observed passing a point during a time interval.

When the involved ingress and egress endnodes are configured to perform such tests, a MEP at the respective network layer, eg LSP tunnel, PW, inserts packets with MPLS-TP test information with specified throughput, packet size and transmission data patterns. For the one way test, the remote MEP receives the packet and calculates the packet loss. For the two way test, the remote MEP loops the packet back to the originating MEP, which calculates the packet loss. The configuration can be done via the management plane or via the control plane. The definition how this is achieved is out side the scope of this draft.

The out-of-service MPLS-TP test function is service affecting, as the test function puts the remote MEP associated with the diagnosed entity into a LOCK, to make sure that the all the frames; including any user or client data frames and TST frames, are properly looped back to the ingress MEP. The source MEP configured for the out-of-service test transmits LCK packets in the immediate client (sub-) layer. Once the LCK is acknowledged, the source MEP gradually increase TST packet bandwidth either via increasing the transmit rate or via increasing frame size, until hitting a preconfigured/defined threshold TST packet traffic loss rate.

For the in-service MPLS-TP test function the user data traffic is not disrupted and the MPLS-TP test packets are transmitted such that a only limited part of the service bandwidth is utilized. The rate and

QoS of transmission for TST packets is pre-determined for in-service MPLS-TP test function. The maximum rate at which TST packets can be sent without adversely impacting the data traffic for an in-service is should be calculated carefully.

Observe TST packet that are transmitted, delivered, and or rejected on a PW, LSP or Section. When detect threshold of packet loss rate, calculated the throughput.

Note: need to explicitly indicate that the test is between two MEPs and that testing is only done between those two points.

Editor's Note TST in service will be updated in next version.

In order to support TST, a Test TLV in TST PDU should be defined:

Test TLV - Optional element whose length and contents are configurable at the MEP. The contents can be a test pattern and an optional checksum. Examples of test patterns include pseudo-random bit sequence (PRBS) (231-1) as specified in sub-clause 5.8 of ITU-T O.150 [4], all '0' pattern, etc.

At the transmitting MEP, provisioning is required for a test signal generator which is associated with the MEP. At a receiving MEP, provisioning is required for a test signal detector which is associated with the MEP.

A MIP is transparent to the TST packets and therefore does not require any provisioning to support MPLS-TP test functionality.

A MEP inserts TST packets towards its peer MEPs. The receiving MEP detects these TST packets and performs the intended measurements.

### 3.2. Transmission

A test signal generator connected to a MEP can transmit TST packets as often as the test signal generator configuration. Each TST packet is transmitted with a specific Sequence Number. A different Sequence Number must be used for every TST packet, and no Sequence Number from the same MEP may be repeated within one minute.

When a MEP is configured for an out-of-service test, the MEP also generates LCK packets in the same direction where TST packets are transmitted. And TST packet transmission rate should be increased gradually by step of x Kb/s and recorded TST packet transmitted, delivered or rejected. This is shown in Figure 1:

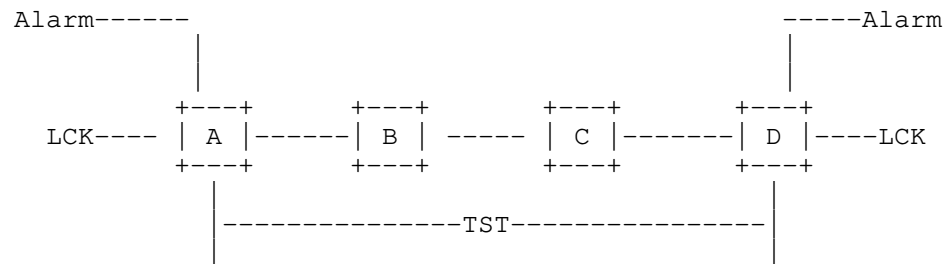


Figure 1: out of service test

LCK is configured by management, When an administrative/diagnostic lock is applied on a MEG, the related MEPs continues to periodically (once per 1s) transmit packets with LCK information until the administrative/diagnostic condition is removed. This allows a client MEP receiving packets with LCK information to differentiate between traffic interruption due to a defect condition and an administrative locking action at the server (sub-) layer MEP.

In case of testing protection path status when it is used in protection switch, QoS of TST packet is setup as same as packet in work path.

When a MEP is configured for an in-service test, the MEP not generates any LCK packet. And TST packet transmission rate should be increased gradually by step of x Kb/s, but it is less than Maximum bit rate. In order to verify the throughput, QoS of test packet should be considered, color, CIR/EIR should be carefully calculated in order not to impact the service.

Editor's note: Details will be updated.

And service packet that is transmitted MUST be also recorded by traffic condition performance counter.

### 3.3. Receive

If the receiving MEP is configured for MPLS-TP test function, the test signal detector connected to the MEP detects bit errors or packet loss rate from e.g. the pseudo-random bit sequence of the received TST packets and reports such errors.

Further, when the receiving MEP is configured for an out-of-service test, it also generates LCK packets a in the direction where the TST packets are received. Detected the packet loss rates or bit errors of test packet, and record the rate of test packet transmission or rejected.

When the receiving MEP is configured for an in service test, no any LCK packet is generated. At same time, record all service packet counters of transmitted, delivered, and or rejected.

### 3.4. Performance Monitoring counter and throughput calculation

To be added.

## 4. TST Frame format (PDU)

TST PDUs are encapsulated by using the ACH, according to RFC 5586 [5].

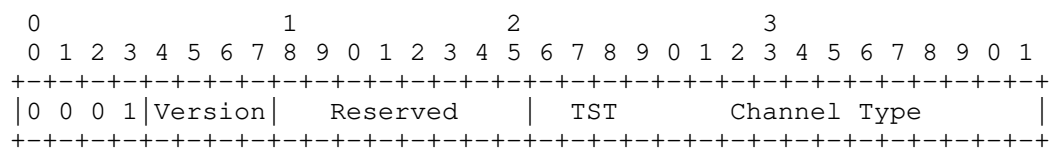


Figure 1: TST-ACH

The first four bytes represent the ACH ([RFC 5586]):

0001: Indicate it is ACH

Version: 00x0

Reserved: reversed for further standardization, it is 00xx

TST Channel type: indicate it is test OAM packet allocated by IANA.

Tools TST use TST PDU to verify bandwidth that carries some information of TST TLV.

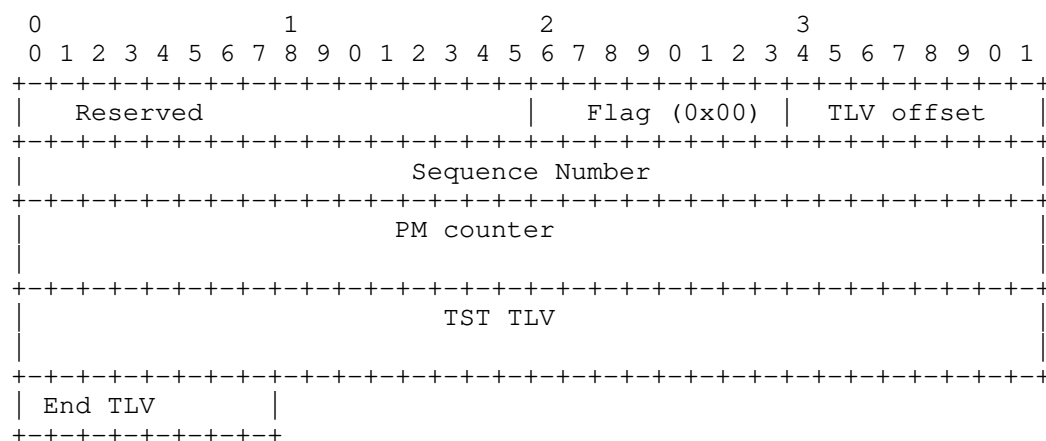


Figure 2: TST PDU

The fields of the TST PDU format are as follows:

Reserved: 16 bits, reserved for future international standardization, set to 00xx.

Flags: none, set to 0x00.

TLV Offset: set to 0x08

Sequence number: 4 octets

PM counter: record packet transmitted, delivered or rejected.

Test TLV: to be inserted in this field, format sees below.



## 5. Security Considerations

Refer to draft-fang-mpls-tp-security-framework [6]

Mechanisms SHOULD be provided to ensure that unauthorized access is prevented from triggering any TST function.

This will prevent unauthorized access to vital equipment and it will prevent third parties from learning about sensitive information about the transport network.

TST messages MAY be authenticated.

## 6. IANA Considerations

There is one channel type for TST by IANA actions required by this draft.

## 7. Acknowledgments

The authors acknowledge the helpful inputs from Xiaobo YI and Italo busi, William Zhang and discussions with Xiaohua MA and Stephan ROULLOT.

## 8. References

### 8.1. Normative References

- [1] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997
- [2] B. Niven-Jenkins, D. Brungard, M. Betts, N. Sprecher, S. Ueno, MPLS-TP Requirements, draft-ietf-mpls-tp-requirements
- [3] M. Vigoureux, D. Ward, M. Betts, Requirements for OAM in MPLS Transport Networks, draft-ietf-mpls-tp-oam-requirements
- [4] ITU-T O.150, General requirements for instrumentation for performance measurements on digital transmission equipment
- [5] M. Bocci, M. Vigoureux, S. Bryant, MPLS Generic Associated Channel, RFC 5586, June 2009

### 8.2. Informative References

- [6] Luyuan Fang, Ben Niven-Jenkins, MPLS-TP security framework, draft-fang-mpls-tp-security-framework

## Authors' Addresses

Feng Huang  
Alcatel-Lucent shanghai Bell  
Email: feng.f.huang@alcatel-sbell.com.cn

Lieven Levrau  
Alcatel-Lucent  
Email: Lieven.Levrau@alcatel-lucent.com

Han Li  
China Mobile  
Email : lihan@chinamobile.com

Ruiquan Jing  
China Telecom  
Email: jingrq@ctbri.com.cn

Network Working Group  
Internet Draft  
Intended status: Standard Track

J.He  
Huawei Technologies

H.Li  
China Mobile

E. Bellagamba  
Ericsson

Expires: January 2011

July 12, 2010

Indication of Client Failure in MPLS-TP  
draft-he-mpls-tp-csf-02.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 January 12, 2011.

Abstract

This document describes a Multi-Protocol Label Switching Transport Profile (MPLS-TP) Operations, Administration and Maintenance (OAM) tool to propagate a client failure indication across an MPLS-TP network in case the propagation of failure status in the client layer is not supported.

## Conventions used in 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 [RFC2119].

## Table of Contents

1. Introduction.....	2
2. Terminology.....	3
3. Mechanisms of CSF.....	3
3.1. General.....	3
3.2. Transmission of CSF.....	5
3.3. Reception of CSF.....	5
3.4. Configuration of CSF.....	5
4. Frame format of CSF.....	6
5. Consequent actions.....	7
6. Security Considerations.....	7
7. IANA Considerations.....	8
8. Acknowledgments.....	8
9. References.....	8
9.1. Normative References.....	8
9.2. Informative References.....	8
10. Authors' Addresses.....	9

## 1. Introduction

In transport network OAM functionalities are important and fundamental to ease operational complexity, enhance network availability and meet service performance objectives by efficient and automatic detection, handling, diagnosis and appropriate reporting of defects and performance monitoring.

In the case of server layer defects detected in a transport network, normally an AIS/FDI is generated for the downstream client signal as an indication to the downstream network elements that the Client signal is missing due to a server layer defect.

According to [MPLS-TP Framework], MPLS-TP clients include PW and network layer clients. Examples of network layer clients include IP, MPLS and MPLS-TP.

In cases the client service to be carried by MPLS-TP networks does not provide mechanisms to propagate its failure information on top of MPLS-TP networks (e.g. not needed in the original application of the client signal, the signal was originally at the bottom of the layer

stack and it was not expected to be transported over a server layer), while such an indication is needed by the downstream, it is necessary that MPLS-TP OAM provides such a tool to help propagate client failure indication to the far end on detection of a failure of the ingress client signal.

This document defines a MPLS-TP OAM tool as Client Signal Fail indication (CSF) to propagate client failures and their clearance across a MPLS-TP domain.

## 2. 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 [RFC2119].

The reader is assumed to be familiar with the terminology in MPLS-TP. The relationship between ITU-T and IETF terminologies on MPLS-TP can be found in [Rosetta stone].

ACH: Associated Channel Header

AIS: Alarm Indication Signal

CSF: Client Signal Fail indication

FDI: Forward Defect Indication

LSR: Label Switching Router

MEP: Maintenance Entity Group End Point

MIP: Maintenance Entity Group Intermediate Point

OAM: Operations, Administration and Maintenance

MPLS-TP: MPLS Transport Profile

RDI: Remote Defect Indication

## 3. Mechanisms of CSF

### 3.1. General

Client Signal Fail indication (CSF) provides a function to enable a MEP to propagate a client failure indication to its peer MEP across a

MPLS-TP network in case the client service itself does not support propagation of its failure status.

Packets with CSF information can be issued by a MEP, upon receiving failure information from its client service. Detection rules for client failure events are client-specific and are therefore outside the scope of this document.

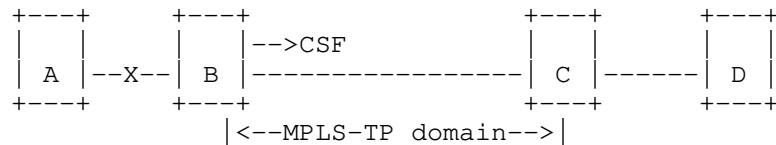


Figure 1 Use case of CSF

Figure 1 depicts a typical connection scenario between two client network elements (Node A and Node D) interconnected through MPLS-TP transport network. Client Node A connects to MPLS-TP Node B and Client Node D connects to MPLS-TP Node C. Node B and C support MPLS-TP MEP function.

If a failure is detected between Node A and Node B and is taken as a native client failure condition, the MEP function in Node B will initiate CSF signal and it will be sent to Node C through MPLS-TP network. CSF signal will be extracted at Node C as an indication of client signal failure. Further, this may be mapped back into native client failure indication and regenerated towards client Node D.

Node B learns the failure between A and B either by direct detection of signal fail (e.g. loss of signal) or by some fault indications between A and B (e.g. RDI, AIS/FDI).

If the connection between Node A and B recovers, Node B may stop sending CSF signals to Node C (implicit failure clearance mechanism) or explicitly send failure clearance indication (e.g. by flags in CSF PDU format) to Node C to help expedite clearance of native client failure conditions.

Accordingly, Node C will clear client failure condition when a valid client data frame is received and no CSF is received (implicit failure clearance mechanism) or upon receiving explicit failure clearance indication.

### 3.2. Transmission of CSF

Upon learning signal failure condition of its client-layer the MEP can immediately start transmitting periodic packets with CSF information. A MEP continues to transmit periodic packets with CSF information until the client-layer signal failure condition is cleared.

The clearance of CSF condition can be communicated to the peer MEP via:

- stopping transmission of CSF signal or
- forwarding CSF PDU with clearance indication.

Transmission of packets with CSF information can be enabled or disabled on a MEP.

Detection and clearance rules for CSF events are client and application specific and outside the scope of this draft.

The period of CSF generation is client and application specific.

### 3.3. Reception of CSF

Upon receiving a packet with CSF information a MEP either declares or clears a client-layer signal fail condition according to the received CSF information and propagates this as a signal fail indication to its client-layer.

### 3.4. Configuration of CSF

Specific configuration information required by a MEP to support CSF transmission is the following:

CSF transmission period - this is application dependent.

PHB - identifies the per-hop behavior of packet with CSF information.

A MIP is transparent to packets with CSF information and therefore does not require any information to support CSF functionality.

#### 4. Frame format of CSF

Figure 2 depicts the frame format of CSF. CSF PDUs are encapsulated using the ACH, according to [RFC 5586].

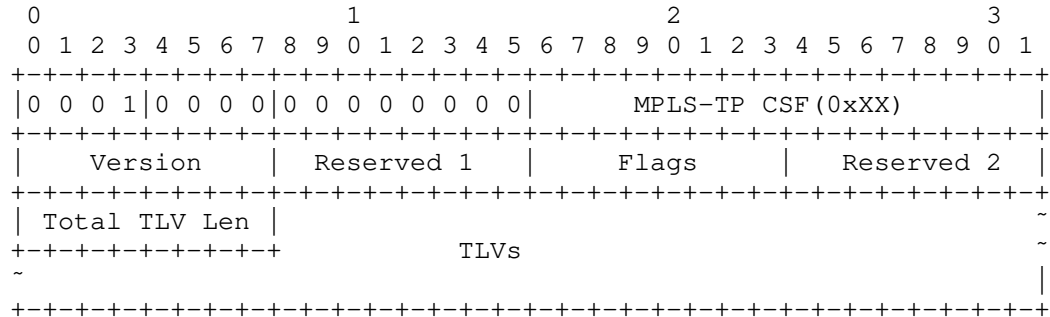


Figure 2 Frame format of CSF

The first four bytes represent the G-ACH ([RFC 5586]):

- first nibble: set to 0001b to indicate a control channel associated with a PW, a LSP or a Section;
- G-ACH Version (bits 4 to 7): set to 0, as specified in [RFC 5586]
- G-ACH Reserved (bits 8 to 15): set to 0 and ignored on reception, as specified in [RFC 5586];
- G-ACH Channel Type (Bits 16 to 31): value 0xXX identifies the payload as CSF PDU. To be assigned by IANA.
- CSF Version (Bits 32 to 39): Set to 0;
- CSF Reserved 1 (Bits 40 to 47): This field MUST be set to zero on transmission and ignored on receipt;
- CSF Reserved 2 (Bits 56 to 63): This field MUST be set to zero on transmission and ignored on receipt;
- Total TLV Length: Total of all included TLVs. No TLVs are defined currently. The value is 0.

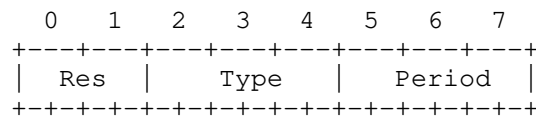


Figure 3 Format of Flags in CSF PDU

Figure 3 depicts the format of Flags in CSF PDU

- Flag Reserved (Bits 48 to 49): Set to 0;
- Type (Bits 50 to 52): Set to the following values to indicate CSF types

Value	Type
000	Client Signal Fail - Loss of Signal (CSF-LOS)
001	Client Signal Fail - Forward Defect Indication (CSF-FDI)
010	Client Signal Fail - Reverse Defect Indication (CSF-RDI)
011	Clearance of Client Signal Fail - (CSF-Clear)

- Period (Bits 53 to 55): CSF transmission period and can be configured.

## 5. Consequent actions

The original usage of CSF is to transport a client signal fail condition at the input of the transport network to the output port of the transport network for clients that do not have AIS defined.

CSF allows the transport network to create a condition at the output port of the transport network such that the customer input port is able to detect and alarm that there is no data arriving i.e. the connection is interrupted. In this case, customers may choose another transport network or another port to continue communication.

## 6. Security Considerations

Malicious insertion of spurious CSF signals (e.g. DoS) is not quite likely in a transport network since transport networks are usually self-managed by operators and providers.

## 7. IANA Considerations

This document requests that IANA allocates a channel type of G-ACH for CSF function to be used in MPLS-TP OAM.

## 8. Acknowledgments

To be added in a future version of the document

## 9. References

### 9.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC5586] Vigoureux, M., Bocci, M., Swallow, G., Ward, D., Aggarwal, R., "MPLS Generic Associated Channel", RFC5586, June 2009
- [ITU-T Recommendation G.7041] "Generic framing procedure (GFP)", ITU-T G.7041, October 2008
- [RFC 5654] Niven-Jenkins, B., Brungard, D., Betts, M., "Requirements of an MPLS Transport Profile", RFC 5654, September 2009
- [RFC 5860] Vigoureux, M., Ward, D., and Betts, M., "Requirements for OAM in MPLS Transport Networks", RFC5860, May 2010
- [RFC 5921] Bocci, M., Bryant, S., Frost, D., "A Framework for MPLS in Transport Networks", RFC 5921, 2010

### 9.2. Informative References

- [MPLS-TP OAM Frmk] Busi, I., Niven-Jenkins, B., Allan, D., "MPLS-TP OAM Framework and Overview", draft-ietf-mpls-tp-oam-framework-06 (work in progress), April 2010
- [Rosetta stone] Van Helvoort, H., Andersson, L., Sprecher, N., "A Thesaurus for the Terminology used in Multiprotocol Label Switching Transport Profile (MPLS-TP) drafts/RFCs and ITU-T's Transport Network Recommendations", draft-ietf-mpls-tp-rosetta-stone-02 (work in progress), May 2010

## 10. Authors' Addresses

Jia He  
Huawei Technologies Co., Ltd.  
  
Email: hejia@huawei.com

Han Li  
China Mobile Communications Corporation  
  
Email: lihan@chinamobile.com

Elisa Bellagamba  
Ericsson  
  
Email: elisa.bellagamba@ericsson.com

## Intellectual Property

The IETF Trust takes no position regarding the validity or scope of any Intellectual Property Rights or other rights that might be claimed to pertain to the implementation or use of the technology described in any IETF Document or the extent to which any license under such rights might or might not be available; nor does it represent that it has made any independent effort to identify any such rights.

Copies of Intellectual Property disclosures made to the IETF Secretariat and any assurances of licenses to be made available, or the result of an attempt made to obtain a general license or permission for the use of such proprietary rights by implementers or users of this specification can be obtained from the IETF on-line IPR repository at <http://www.ietf.org/ipr>

The IETF invites any interested party to bring to its attention any copyrights, patents or patent applications, or other proprietary rights that may cover technology that may be required to implement any standard or specification contained in an IETF Document. Please address the information to the IETF at [ietf-ipr@ietf.org](mailto:ietf-ipr@ietf.org).

The definitive version of an IETF Document is that published by, or under the auspices of, the IETF. Versions of IETF Documents that are published by third parties, including those that are translated into other languages, should not be considered to be definitive versions of IETF Documents. The definitive version of these Legal Provisions is that published by, or under the auspices of, the IETF. Versions of these Legal Provisions that are published by third parties, including those that are translated into other languages, should not be considered to be definitive versions of these Legal Provisions.

For the avoidance of doubt, each Contributor to the IETF Standards Process licenses each Contribution that he or she makes as part of the IETF Standards Process to the IETF Trust pursuant to the provisions of RFC 5378. No language to the contrary, or terms, conditions or rights that differ from or are inconsistent with the rights and licenses granted under RFC 5378, shall have any effect and shall be null and void, whether published or posted by such Contributor, or included with or in such Contribution.

#### Disclaimer of Validity

All IETF Documents and the information contained therein are provided on an "AS IS" basis and THE CONTRIBUTOR, THE ORGANIZATION HE/SHE REPRESENTS OR IS SPONSORED BY (IF ANY), THE INTERNET SOCIETY, THE IETF TRUST AND THE INTERNET ENGINEERING TASK FORCE DISCLAIM ALL WARRANTIES, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO ANY WARRANTY THAT THE USE OF THE INFORMATION THEREIN WILL NOT INFRINGE ANY RIGHTS OR ANY IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE.

#### 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.



MPLS Working Group  
Internet-Draft  
Intended Status: Standards Track  
Expires: April 2011

Y. Huang  
Y. Jiang  
Huawei Technologies  
October 25, 2010

Signaling extension for MPLS ring protection  
draft-huang-mpls-ring-signaling-extension-01.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 April 25, 2011.

Abstract

When using Multi-Protocol Label Switching (MPLS) Label Switched Paths (LSP) on a ring topology, it is needed to minimize the labels for protection purpose and minimize the configuration effort. Though MPLS Fast Re-Route [MPLS-FRR] can achieve automatic LSP service protection, it is not optimized as many labels are used to create lots of detour LSPs.

This document describes a MPLS ring mechanism and some extensions on signaling protocol to facilitate the establishment of Label Switched Paths (LSPs) using Label Distribution Protocol [LDP] or RSVP-TE protocol [RSVP-TE] to achieve automated link and node protection.

## Table of Contents

1. Introduction.....	2
2. Conventions used in this document.....	3
3. Ring Protection Scheme.....	3
3.1. P-to-p LSP example.....	4
3.1.1. Link Failure example.....	4
3.1.2. Node Failure example.....	6
3.2. p-t-mp LSP example.....	7
3.2.1. Link Failure example.....	7
3.2.2. Node Failure example.....	7
3.3. OAM Consideration.....	7
4. Signaling extension.....	8
4.1. LDP extension.....	10
4.2. RSVP-TE extension.....	12
5. Security Considerations.....	12
6. IANA Considerations.....	12
7. Conclusions.....	12
8. References.....	12
8.1. Normative References.....	12
8.2. Informative References.....	13
9. Acknowledgments.....	13

## 1. Introduction

More and more network operators have considered using unified IP/MPLS technology in a single network infrastructure to provide multi-service, including fixed and mobile services. Since a large amount of access and aggregation network segments are fiber ring topology network, it is expected to use MPLS on a ring topology network. The industry is interested to find a lightweight planning, easy provision and little resource consumption solution for MPLS ring.

MPLS Fast Re-Route [MPLS-FRR] can automatically create protection LSP and minimize the network provision work, but it is not optimized since it uses many labels and creates lots of detour LSPs, especially in ring topology.

This document describes a simple MPLS ring mechanism under which numerous service LSPs can be created automatically across the ring with some extensions on signaling protocol. The extensions on signaling solves label switch disruption because of node failure. Section 3 describes the ring protection scheme based on wrapping

mechanism. Section 4 describes the signaling extensions for both LDP and RSVP-TE.

## 2. Conventions used in this document

In examples, "C:" and "S:" indicate lines sent by the client and server respectively.

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].

CW            Clock Wise

CCW          Counter Clock Wise

MPLS        Multi-Protocol Label Switching

MPLS-TP    MPLS Transport Profile

SPME        Sub-path Maintenance Element

## 3. Ring Protection Scheme

Several Internet Drafts on MPLS Ring said that both Steering mode and Wrapping mode can be used in a ring network. The scheme in this document only discusses wrapping mechanism. The idea is that steering method takes no advantage of ring topology and can just be achieved by known technology, such as MPLS TE FRR or PW redundancy. It is not intended to compare the two methods and to say which one is better than the other in this document and it should only be the operator's choices.

In general, the scheme includes several technical points as following.

1. For data transport layer on the ring, for each upstream /downstream direction, assigning CW as working path direction and CCW as protection path direction or vice versa. (The following examples in this document take CW as working direction on the ring). On working ring direction (CW), the packet outermost label is service LSP label. That is to say, no SPME label be added to service LSP packet at ring working direction.

2. Pre-provision a closed protection LSP on ring protection direction. The ring protection LSP is to protect all service LSPs when a link or node failure occurs.

3. Control layer, service LSPs crossing the ring can be created normally by LSP signaling with the extension in section 4. The extension is to send label switch information to downstream ring node or backup node.

4. Normally, service LSP packet is forwarded along working direction, each ring node performs the normal MPLS forwarding at service LSP label level.

5. When a link or node failure occurs, upstream node adjacent to the failure performs wrapping and pushes a protection ring label to the service LSP packet, and the packet is wrapped around in the protection ring direction and goes to downstream node adjacent to the failure.

6. For a link failure, the downstream node adjacent to the failure is the immediate downstream node, it just pops the protection label and forwards the packet to the working direction.

7. For a node failure, the downstream node adjacent to the failure is one hop away from the upstream node adjacent to the failure. After popping the protection label, the downstream node adjacent to the failure can do the label switching work in working direction with additional label switching information got from the failure node during the setup of the service LSP. The detailed procedure and the definition of the information is referred to section 4.

8. To protect failure of ingress or egress node of a ring, a backup node can be assigned during MPLS ring provisioning. The signaling extensions defined in section 4 also provide the ability to notify the backup node the label switching information.

Following sections provide some examples in details.

### 3.1. P-to-p LSP example

#### 3.1.1. Link Failure example

A LSP (LSP 1 in figure 1) crosses ring nodes A->B->C->D with labels: . . . [L1]->[L2]->[L3]->[L4]->[L5]-> . . .

A protection LSP is a closed LSP in CCW direction and is denoted as: [p1]->[p2]->[p3]->[p4]->[p5]->[p6]->[p1].

Here we give some symbol illustration for the label stack

1. The label stack will be enclosed in square brackets ("[]")
2. Each level in the stack will be separated by the '|' character

3. The bottom of the stack will be denoted by the string "B" for two or more label layer.

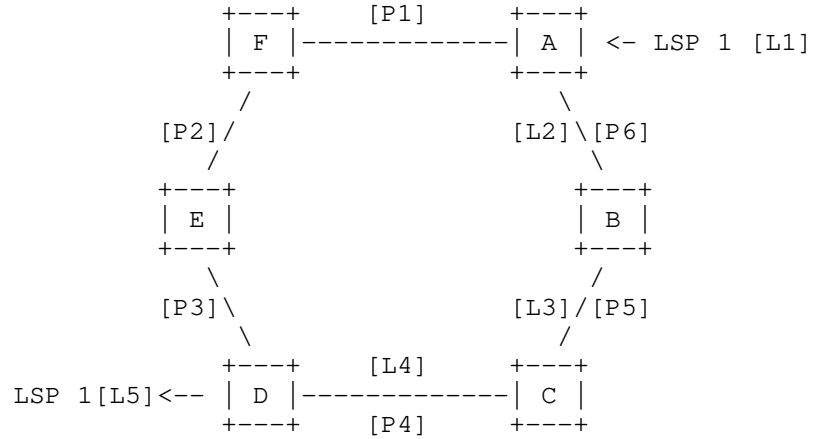


Figure 1: Labels allocation example for p-t-p LSP protection

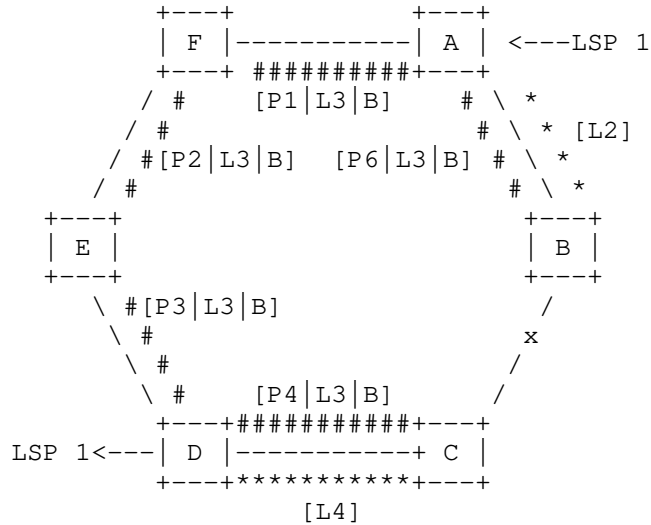


Figure 2: packet flow when link B-C failure

For example, the link between B and C goes down as shown in figure 2. After recovering, the packet forwarding path is denoted by '\*\*\*\*' in the working direction and '###' in the protection direction.

The forwarding path and encapsulation on link is: A-> [L2]-> B -> [P6|L3|B] -> A -> [P1|L3|B] -> F -> [P2|L3|B] -> E -> [P3|L3|B] -> D -> [P4|L3|B] -> C -> [L4] -> D.

When node B finds the link to C is failed, it push protection label [P5] to the out going packet with [L3] at egress port to C, so the label stack is [P5|L3|B], do wrapping and switch [P5|L3|B] to [P6|L3|B], then send out to link B-A.

Node A,F,E,D just do Protection label switching and send the packet to C, and C receives the packet with [P4|L3|B].

When node C finds the link to B has failed, at egress port to B, it pops protection label [P5] and extracts [L3], do wrapping and switch [L3] to [L4] based on normal label switching table entry, then send out [L4] to link C-D. Then at node D, the service LSP drops the ring after normal label switching.

### 3.1.2. Node Failure example

For example, the Node B goes down as shown in figure 3.

After recovering, the packet forwarding path is denoted by '\*\*\*\*' in working direction and '###' in protection direction.

The forwarding path and the encapsulation on link is: A -> [P1|L2|B] -> F -> [P2|L2|B] -> E -> [P3|L2|B] -> D -> [P4|L2|B] -> C -> [L4] -> D.

When node A finds the link to B is failed, it push protection label [P6] to the outgoing packet with [L2] at egress port to B, thus the label stack is [P6|L2|B], do wrapping and switch [P6|L2|B] to [P1|L2|B], then send out to link A-F.

Node F,E,D just do Protection label switching and send the packet to C, and C receives the packet with [P4|L2|B].

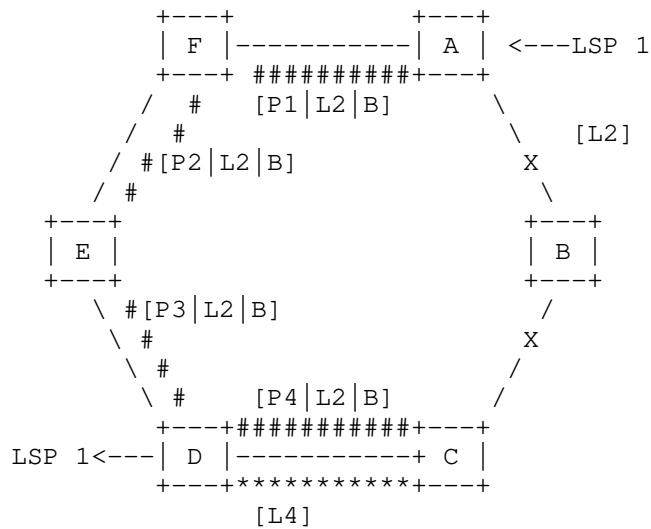


Figure 3: packet flow when node B failure

When node C finds the link to B is failed, it pops the protection label [P5] and gets [L2] at egress port to B, does wrapping and switches [L2] to [L4] supported by label switching information notified by node B during setup of the LSP. This notification operation uses the signaling extensions defined in section 4. Node C then sends out packet with [L4] to link C-D. At node D, the service LSP drops the ring after normal label switching.

### 3.2. p-t-mp LSP example

TBD

#### 3.2.1. Link Failure example

TBD

#### 3.2.2. Node Failure example

TBD

### 3.3. OAM Consideration

Each pair of adjacent nodes on the ring should deploy OAM continuity check (CC) mechanism to detecting failures between each other. When one node gets event(s) that the CC check to one peer node fails, it will inform from the opposite direction to inform the event for the ring. That is to say, using section OAM can do the job.

Section OAM also can meet the need for multiple logical ring instances be deployed in one physical link. The possible scenario in industry regarding this is two tangent rings which share one or more links between two ring nodes (the ring nodes can be adjacent or not). Theoretically one may worry about the case that when do wrapping, how to recognize different logical ring and push related protection label on the packets. In reality, as long as we design the two ring under the same policy, that is, taking CW (or vice versa) as working direction on both rings (this is the most popular case), there is no problem in question on the shared link(s).

How to perform the OAM CC and how to inform the failure event to the right receiver, is out of scope of this document. We think that both MPLS and MPLS-TP OAM mechanism can be used for the purpose.

#### 4. Signaling extension

The purpose of the signaling extension is to backup the protected node's label switching information to one backup node. Under ring topology the backup type has two case:

Case 1: For protecting transit ring node, the backup node is the immediate downstream node to the protected one.

Case 2: For protecting ingress/egress ring node, backup node is provisioned by operator.

Case 1 can be illustrated in the same example shown in figure 3, the normal forwarding path and labels on links are: A -> [L2] -> B -> [L3] -> C -> [L4] -> D. During the LSP setup, when node B gets the mapping label from C [L3], it allocate [L2] to node A and locally generate label switching information <L2-to-L3> for forwarding. At this point of time, Node C send a new message to Node C containing the local switching information <L2-to-L3>.

When node C gets the message containing the label mapping information <L2-to-L3>, combines it's local label switching information <L3-to-L4>, and generates a new label mapping info as <L2-to-L4>. The new label switching information can help node C to finish the work described in section 3.1.2.

Every protected node should perform the signaling procedure like node B in above example.

Note that above mechanism needs to do some label planning work to avoid duplicate label for adjacent links (e.g. link C-D and link B-C are adjacent links).

Case 2 can be illustrated in figure 4.

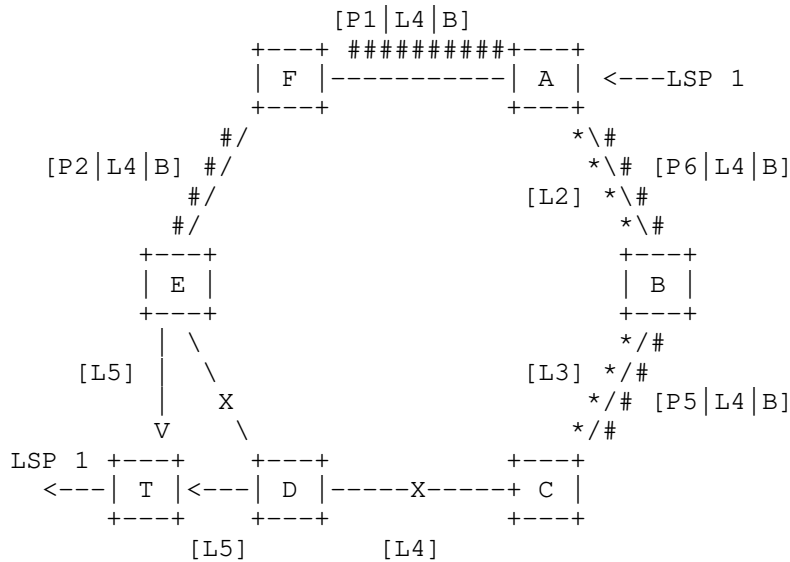


Figure 4 Packet flow when node D failure

Definition for illustration:

Port E-T : Means physical port at node E which connects node T.

Backup id: An id number to correlate two ports.

To protect node D which is the egress node of some service LSPs (e.g. LSP 1 in figure 4), a backup node E should be selected and configured. The port E-T and port D-T are configured one same backup id for example 1000.

When node D gets the allocated label [L5] from downstream node T which is not the ring node, it allocate label [L4] to node C. At the mean time, node D send a new message containing label mapping information <L4-to-L5> and backup id 1000 to pre-configured backup node E.

At node E, after getting the message, E generate a new local label switching information <L4-to-L5> take in-port as ring port E-D and out-port as the port E-T according to backup id.

In figure 4, When E gets packets from the protection ring direction with encapsulation [P2|L4|B], it pop the protection label and switch [L4] to [15] and send out the packet to node T. Node T can handle the packet from port T-E just like it comes from port T-D.

#### 4.1. LDP extension

A TLV bearing the label mapping information described above can meet both case 1 and case 2. The TLV can be named as Label Mapping Information TLV.

For LDP protocol, it use Notification message ([LDP] Section 3.5.1 ) to contain the new TLV. The TLV is optional.

If a LSR is configured to perform ring protection as section 3, after it gets one label (egress label) from it's downstream peer and allocate a new label (ingress label) to the peer LSR which stands at upstream, it SHOULD send a Notification message containing the Label Mapping Information TLV to the downstream LSR under the Case 1 or to the backup LSR under the Case 2. The LSR can discriminate the different cases by the fact that whether the egress port for the label is the port which belongs to the ring.

If a LSR gets the Notification message which contains the Label-Mapping-Info TLV, it SHOULD generate a new local label switching information for forwarding. The LSR can recognize the cases by the TLV content.

The Label-Mapping-Info TLV will be put in the optional parameter field following the status TLV and be described as following:

Optional Parameter	Type	Length	Value
Label-Mapping-Info	TBD	20	See below

Label Mapping Information TLV 's format is shown in figure 5.

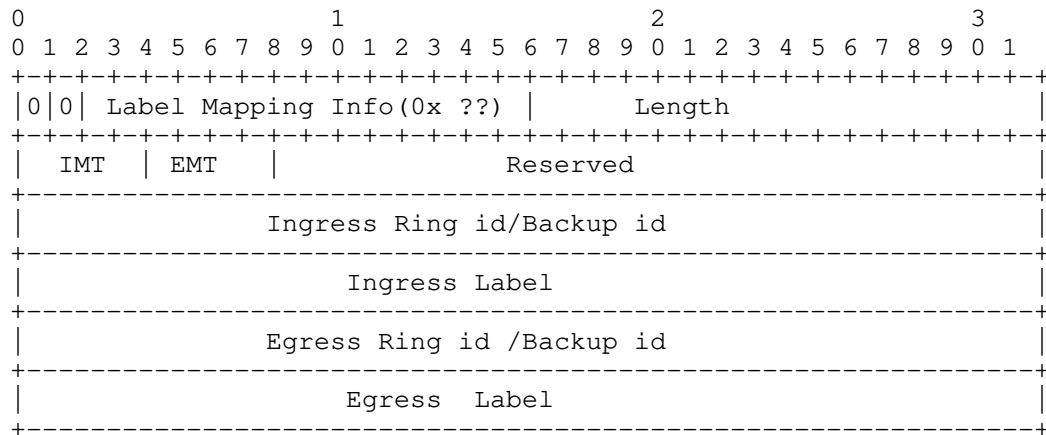


Figure 5 Label Mapping Info TLV format in LDP

Following gives some explanation of Label Mapping Info TLV.

IMT Ingress Mapping Type

0: Denote that the value in field Ingress Ring id/Backup id is the ring id to which the LSP belongs at the ingress port.

1: Denote that the value in field Ingress Ring id/Backup id is the backup id which the ingress port correlates.

Other: Reserved.

EMT Egress Mapping Type

0: Denote that the value in field Egress Ring id/Backup id is the ring id to which the LSP belongs at the egress port.

1: Denote that the value in field Egress Ring id/Backup id is the backup id which the egress port correlates.

Other: Reserved.

Ingress ring/backup id: The value of ingress ring id or backup id of the ingress port.

Ingress Label: The value of ingress label.

Egress ring/backup id: The value of egress ring id or backup id of the egress port.

Egress Label: The value of Egress label.

#### 4.2. RSVP-TE extension

TBD

#### 5. Security Considerations

TBD

#### 6. IANA Considerations

A new LDP protocol TLV code for Label-Mapping-Info TLV need to be assigned by IANA.

RSVP-TE extension: TBD.

#### 7. Conclusions

TBD

#### 8. References

##### 8.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [MPLS-FRR] P. Pan, G. Swallow and A. Atlas., " Fast Reroute Extensions to RSVP-TE for LSP Tunnels ", BCP 14, RFC 4090, May 2005.
- [LDP] Andersson, L., Ed., Minei, I., Ed., and B. Thomas, Ed., "LDP Specification", RFC 5036, October 2007.
- [RSVP-TE] Awduche, D., et al, "RSVP-TE: Extensions to RSVP for LSP Tunnels", RFC 3209, December 2001.

## 8.2. Informative References

- [Inf-1] Y. Weingarten, et al, "MPLS-TP Ring Protection", August 2010.
- [Inf-2] I. Umansky, et al, "MPLS-TP Ring Protection Switching (MRPS)", August 2010.
- [Inf-3] S. Kini, et al, "Efficient Fast Re-route (FRR) using Facility backup in ring topology", August 2010.

## 9. Acknowledgments

This document was prepared using 2-Word-v2.0.template.dot.

## Authors' Addresses

Yong Huang  
Huawei Technologies Co., Ltd.  
Bantian industry base, Longgang district  
Shenzhen, China  
Email: huang.yong@huawei.com

Yuanlong Jiang  
Huawei Technologies Co., Ltd.  
Bantian industry base, Longgang district  
Shenzhen, China  
Email: yljiang@huawei.com

## 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.



Internet Draft  
Category: Informational  
Expiration Date: April 15, 2011

Loa Andersson, Ed. (Ericsson)  
Lou Berger, Ed. (LabN)  
Luyuan Fang, Ed. (Cisco)  
Nabil Bitar, Ed. (Verizon)  
Eric Gray, Ed. (Ericsson)

October 15, 2010

## MPLS-TP Control Plane Framework

draft-ietf-ccamp-mpls-tp-cp-framework-03.txt

### Abstract

The MPLS Transport Profile (MPLS-TP) supports static provisioning of transport paths via a Network Management System (NMS), and dynamic provisioning of transport paths via a control plane. This document provides the framework for MPLS-TP dynamic provisioning, and covers control plane addressing, routing, path computation, signaling, traffic engineering, and path recovery. MPLS-TP uses GMPLS as the control plane for MPLS-TP LSPs. MPLS-TP also uses the control plane for Pseudowires (PWs). Management plane functions such as manual configuration and the initiation of LSP setup are out of scope of this document.

This document is a product of a joint Internet Engineering Task Force (IETF) / International Telecommunication Union Telecommunication Standardization Sector (ITU-T) effort to include an MPLS Transport Profile within the IETF MPLS and Pseudowire Emulation Edge-to-Edge (PWE3) architectures to support the capabilities and functionalities of a packet transport network as defined by the ITU-T.

This Informational Internet-Draft is aimed at achieving IETF Consensus before publication as an RFC and will be subject to an IETF Last Call.

[RFC Editor, please remove this note before publication as an RFC and insert the correct Streams Boilerplate to indicate that the published RFC has IETF consensus.]

### 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), 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/lid-abstracts.html>

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

This Internet-Draft will expire on April 15, 2011

#### Copyright and License 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
1.1	Scope .....	4
1.2	Basic Approach .....	5
1.3	Reference Model .....	6
2	Control Plane Requirements .....	9
2.1	Primary Requirements .....	9
2.2	MPLS-TP Framework Derived Requirements .....	18
2.3	OAM Framework Derived Requirements .....	19
2.4	Security Requirements .....	24
2.5	Identifier Requirements .....	24
3	Relationship of PWs and TE LSPs .....	25
4	TE LSPs .....	26
4.1	GMPLS Functions and MPLS-TP LSPs .....	26
4.1.1	In-Band and Out-Of-Band Control .....	26
4.1.2	Addressing .....	27
4.1.3	Routing .....	28
4.1.4	TE LSPs and Constraint-Based Path Computation .....	28
4.1.5	Signaling .....	29
4.1.6	Unnumbered Links .....	29

4.1.7	Link Bundling .....	29
4.1.8	Hierarchical LSPs .....	29
4.1.9	LSP Recovery .....	30
4.1.10	Control Plane Reference Points (E-NNI, I-NNI, UNI) .....	31
4.2	OAM, MEP (Hierarchy), MIP Configuration and Control ....	31
4.2.1	Management Plane Support .....	31
4.3	GMPLS and MPLS-TP Requirements Table .....	32
4.4	Anticipated MPLS-TP Related Extensions and Definitions .	36
4.4.1	MPLS-TE to MPLS-TP LSP Control Plane Interworking .....	36
4.4.2	Associated Bidirectional LSPs .....	36
4.4.3	Asymmetric Bandwidth LSPs .....	36
4.4.4	Recovery for P2MP LSPs .....	37
4.4.5	Test Traffic Control and other OAM functions .....	37
4.4.6	DiffServ Object usage in GMPLS .....	37
4.4.7	Support for MPLS-TP LSP Identifiers .....	37
4.4.8	Support for MPLS-TP Maintenance Identifiers .....	38
5	Pseudowires .....	38
5.1	LDP Functions and Pseudowires .....	38
5.2	PW Control (LDP) and MPLS-TP Requirements Table .....	39
5.3	Anticipated MPLS-TP Related Extensions .....	41
5.3.1	Extensions to Support Out-of-Band PW Control .....	42
5.3.2	Support for Explicit Control of PW-to-LSP Binding .....	42
5.3.3	Support for Dynamic Transfer of PW Control/Ownership ...	43
5.3.4	Interoperable Support for PW/LSP Resource Allocation ...	43
5.3.5	Support for PW Protection and PW OAM Configuration ....	44
5.3.6	Client Layer and Cross-Provider Interfaces to PW Control.	45
5.4	ASON Architecture Considerations .....	45
6	Security Considerations .....	45
7	IANA Considerations .....	46
8	Acknowledgments .....	46
9	References .....	46
9.1	Normative References .....	46
9.2	Informative References .....	49
10	Authors' Addresses .....	54

## 1. Introduction

The MPLS Transport Profile (MPLS-TP) is being defined in a joint effort between the International Telecommunications Union (ITU) and the IETF. The requirements for MPLS-TP are defined in the requirements document, see [RFC5654]. These requirements state that "A solution MUST be provided to support dynamic provisioning of MPLS-TP transport paths via a control plane." This document provides the framework for such dynamic provisioning.

This document is a product of a joint Internet Engineering Task Force (IETF) / International Telecommunications Union Telecommunications Standardization Sector (ITU-T) effort to include an MPLS Transport Profile within the IETF MPLS and PWE3 architectures to support the capabilities and functions of a packet transport network as defined by the ITU-T.

## 1.1. Scope

This document covers the control plane functions involved in establishing MPLS-TP Label Switched Paths (LSPs) and Pseudowires (PWs). The control plane requirements for MPLS-TP are defined in the MPLS-TP requirements document [RFC5654]. These requirements define the role of the control plane in MPLS-TP. In particular, Section 2.4 of [RFC5654] and portions of the remainder of Section 2 of [RFC5654] provide specific control plane requirements.

The LSPs provided by MPLS-TP are used as a server layer for IP, MPLS and PWs, as well as other tunneled MPLS-TP LSPs. The PWs are used to carry client signals other than IP or MPLS. The relationship between PWs and MPLS-TP LSPs is exactly the same as between PWs and MPLS LSPs in an MPLS Packet Switched Network (PSN). The PW encapsulation over MPLS-TP LSPs used in MPLS-TP networks is also the same as for PWs over MPLS in an MPLS network. MPLS-TP also defines protection and restoration (or, collectively, recovery) functions, see [RFC5654] and [RFC4427]. The MPLS-TP control plane provides methods to establish, remove and control MPLS-TP LSPs and PWs. This includes control of data plane, OAM and recovery functions.

A general framework for MPLS-TP has been defined in [RFC5921], and a survivability framework for MPLS-TP has been defined in [TP-SURVIVE]. These document scope the approaches and protocols that are the foundation of MPLS-TP. Notably, Section 3.5 of [RFC5921] scopes the IETF protocols that serve as the foundation of the MPLS-TP control plane. The PW control plane is based on the existing PW control plane, see [RFC4447], and the PW end-to-end (PWE3) architecture, see [RFC3985]. The LSP control plane is based on Generalized MPLS (GMPLS), see [RFC3945], which is built on MPLS Traffic Engineering (TE) and its numerous extensions. [TP-SURVIVE] focuses on the recovery functions that must be supported within MPLS-TP. It does not specify which control plane mechanisms are to be used.

The remainder of this document discusses the impact of the MPLS-TP requirements on the GMPLS signaling and routing protocols that are used to control MPLS-TP LSPs, and on the control of PWs as specified in [RFC4447], [SEGMENTED-PW], and [MS-PW-DYNAMIC].

## 1.2. Basic Approach

The basic approach taken in defining the MPLS-TP Control Plane framework is:

- 1) MPLS technology as defined by the IETF is the foundation for the MPLS Transport Profile.
- 2) The data plane for MPLS-TP is a standard MPLS data plane [RFC3031] as profiled in [RFC5960].
- 3) MPLS PWs are used by MPLS-TP including the use of targeted LDP as the foundation for PW signaling [RFC4447]; and OSPF-TE, ISIS-TE or MP-BGP as they apply for Multi-Segment (MS)-PW routing. However, the PW can be encapsulated over an MPLS-TP LSP (established using methods and procedures for MPLS-TP LSP establishment) in addition to the presently defined methods of carrying PWs over LSP based packet switched networks (PSNs). That is, the MPLS-TP domain is a packet switched network from a PWE3 architecture perspective [RFC3985].
- 4) The MPLS-TP LSP control plane builds on the GMPLS control plane as defined by the IETF for transport LSPs. The protocols within scope are RSVP-TE [RFC3473], OSPF-TE [RFC4203][RFC5392], and ISIS-TE [RFC5307][RFC5316]. ASON signaling and routing requirements in the context of GMPLS can be found in [RFC4139] and [RFC4258].
- 5) Existing IETF MPLS and GMPLS RFCs and evolving Working Group Internet-Drafts should be reused wherever possible.
- 6) If needed, extensions for the MPLS-TP control plane should first be based on the existing and evolving IETF work, secondly based on work by other standard bodies only when IETF decides that the work is out of the IETF's scope. New extensions may be defined otherwise.
- 7) Extensions to the GMPLS control plane may be required in order to fully automate MPLS-TP LSP related functions.
- 8) Control plane software upgrades to existing (G)MPLS enabled equipment is acceptable and expected.
- 9) It is permissible for functions present in the GMPLS and PW control planes to not be used in MPLS-TP networks.
- 10) One possible use of the control plane is to configure, enable and generally control OAM functionality. This will require extensions to existing control plane specifications which will be usable in MPLS-TP as well as MPLS networks.
- 11) The foundation for MPLS-TP control plane requirements is primarily found in Section 2.4 of [RFC5654] and relevant portions of the remainder Section 2 of [RFC5654].

### 1.3. Reference Model

The control plane reference model is based on the general MPLS-TP reference model as defined in the MPLS-TP framework [RFC5921]. Per the MPLS-TP framework [RFC5921], the MPLS-TP control plane is based on GMPLS with RSVP-TE for LSP signaling and targeted LDP for PW signaling. In both cases, OSPF-TE or ISIS-TE with GMPLS extensions is used for dynamic routing within an MPLS-TP domain.

Note that in this context, "targeted LDP" (or T-LDP) means LDP as defined in RFC 5036, using Targeted Hello messages. See Section 2.4.2 ("Extended Discovery Mechanism") of [RFC5036]. Use of the extended discovery mechanism is specified in [RFC4447] Section 5 ("LDP").

From a service perspective, MPLS-TP client services may be supported via both PWs and LSPs. PW client interfaces, or adaptations, are defined on an interface technology basis, e.g., Ethernet over PW [RFC4448]. In the context of MPLS-TP LSP, the client interface is provided at the network layer and may be controlled via a GMPLS based UNI, see [RFC4208], or statically provisioned. As discussed in [RFC5921], MPLS-TP also presumes an LSP NNI reference point.

The MPLS-TP end-to-end control plane reference model is shown in Figure 1. The Figure shows the control plane protocols used by MPLS-TP, as well as the UNI and NNI reference points, in the case of a single segment PW. (The MS-PW case is not shown.)

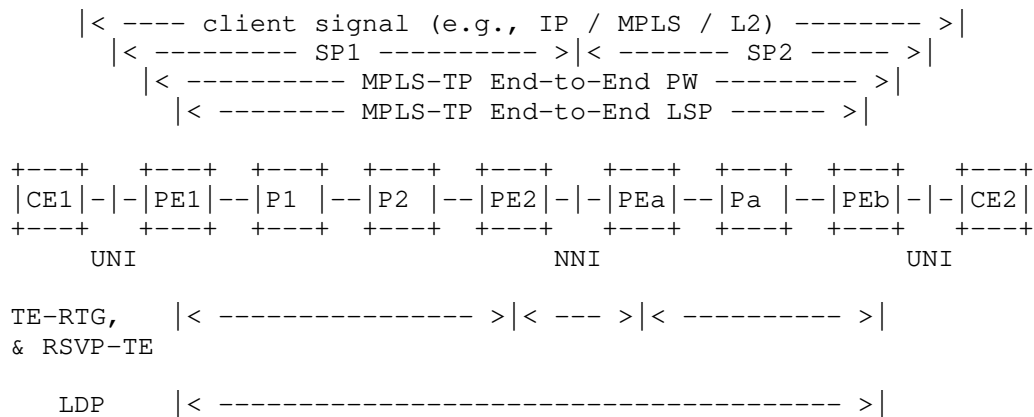


Figure 1. End-to-End MPLS-TP Control Plane Reference Model

Legend:

CE: Customer Edge  
Client signal: defined in MPLS-TP Requirements  
L2: Any layer 2 signal that may be carried over a PW, e.g. Ethernet.  
NNI: Network to Network Interface  
PE: Provider Edge  
SP: Service Provider  
TE-RTG: OSPF-TE or ISIS-TE  
UNI: User to Network Interface

Figure 2 adds three hierarchical LSP segments, labeled as "H-LSPs". These segments are present to support scaling, OAM and Maintenance End Points (MEPs), see [TP-OAM], within each provider domain and across the inter-provider NNI. The MEPs are used to collect performance information, support diagnostic and fault management functions, and support OAM triggered survivability schemes as discussed in [TP-SURVIVE]. Each H-LSP may be protected or restored using any of the schemes discussed in [TP-SURVIVE]. End-to-end monitoring is supported via MEPs at the End-to-End LSP and PW end points. Note that segment MEPs may be collocated with MIPs of the next higher-layer (e.g., end-to-end) LSPs. H-LSPs may also be used to implement Sub-Path Maintenance Elements (SPMEs) as defined in [RFC5921]. (The MS-PW case is not shown.)

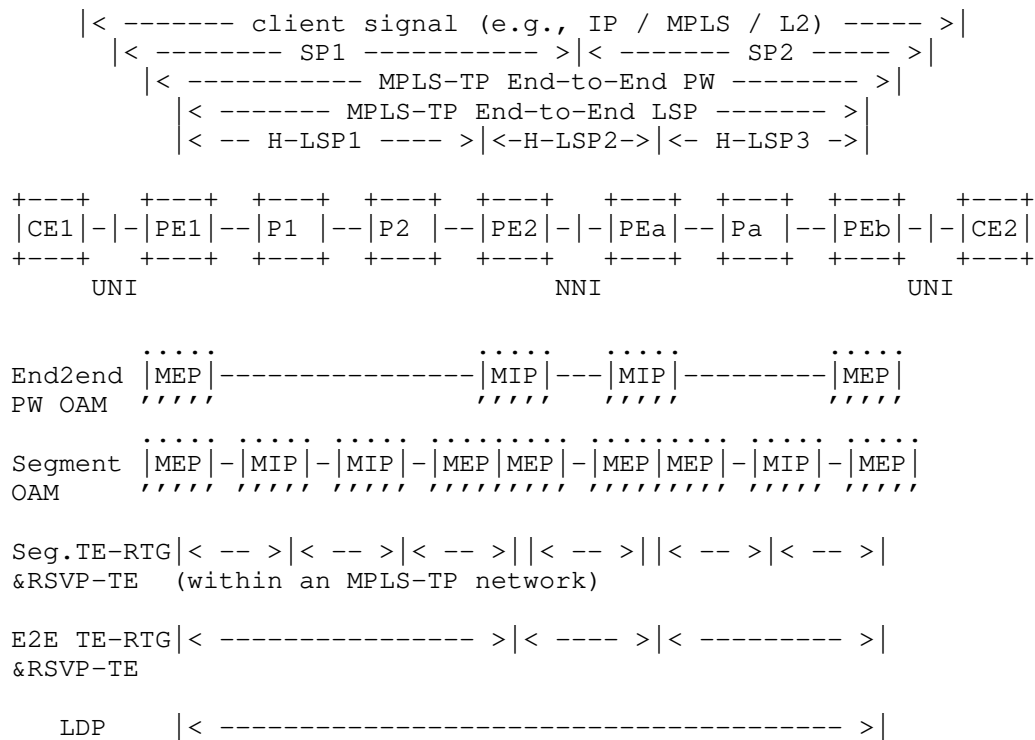


Figure 2. MPLS-TP Control Plane Reference Model with OAM

Legend:

CE: Customer Edge  
Client signal: defined in MPLS-TP Requirements  
E2E: End-to-end  
L2: Any layer 2 signal that may be carried over a PW, e.g. Ethernet.  
H-LSP: Hierarchical LSP  
MEP: Maintenance end point  
MIP: Maintenance intermediate point  
NNI: Network to Network Interface  
PE: Provider Edge  
SP: Service Provider  
TE-RTG: OSPF-TE or ISIS-TE

While not shown in the Figures above, the MPLS-TP control plane must support the addressing separation and independence between the data, control and management planes. Address separation between the planes is already included in GMPLS. Such separation is also already included in LDP as LDP session end point addresses are never automatically associated with forwarding.

## 2. Control Plane Requirements

The requirements for the MPLS-TP control plane are derived from the MPLS-TP requirements and framework documents, specifically [RFC5654], [RFC5921], [RFC5860], [TP-OAM], and [TP-SURVIVE]. The requirements are summarized in this section, but do not replace those documents. If there are differences between this section and those documents, those documents shall be considered authoritative.

### 2.1. Primary Requirements

These requirements are based on Section 2 of [RFC5654]:

1. Any new functionality that is defined to fulfill the requirements for MPLS-TP must be agreed within the IETF through the IETF consensus process as per [RFC4929] [RFC5654, Section 1, Paragraph 15].
2. The MPLS-TP control plane design should as far as reasonably possible reuse existing MPLS standards [RFC5654, requirement 2].
3. The MPLS-TP control plane must be able to interoperate with existing IETF MPLS and PWE3 control planes where appropriate [RFC5654, requirement 3].
4. The MPLS-TP control plane must be sufficiently well-defined to ensure the interworking between equipment supplied by multiple vendors will be possible both within a single domain and between domains [RFC5654, requirement 4].
5. The MPLS-TP control plane must support a connection-oriented packet switching model with traffic engineering capabilities that allow deterministic control of the use of network resources [RFC5654, requirement 5].
6. The MPLS-TP control plane must support traffic-engineered point-to-point (P2P) and point-to-multipoint (P2MP) transport paths [RFC5654, requirement 6].
7. The MPLS-TP control plane must support unidirectional, associated bidirectional and co-routed bidirectional point-to-point transport paths [RFC5654, requirement 7].
8. The MPLS-TP control plane must support unidirectional point-to-multipoint transport paths [RFC5654, requirement 8].
9. The MPLS-TP control plane must enable all nodes (i.e., ingress, egress and intermediate) to be aware about the pairing relationship of the forward and the backward directions belonging to the same co-routed bidirectional transport path

[RFC5654, requirement 10].

10. The MPLS-TP control plane must enable edge nodes (i.e., ingress and egress) to be aware of the pairing relationship of the forward and the backward directions belonging to the same associated bidirectional transport path [RFC5654, requirement 11].
11. The MPLS-TP control plane should enable common transit nodes to be aware of the pairing relationship of the forward and the backward directions belonging to the same associated bidirectional transport path [RFC5654, requirement 12].
12. The MPLS-TP control plane must support bidirectional transport paths with symmetric bandwidth requirements, i.e. the amount of reserved bandwidth is the same in the forward and backward directions [RFC5654, requirement 13].
13. The MPLS-TP control plane must support bidirectional transport paths with asymmetric bandwidth requirements, i.e. the amount of reserved bandwidth differs in the forward and backward directions [RFC5654, requirement 14].
14. The MPLS-TP control plane must support the logical separation of the control plane from the management and data plane [RFC5654, requirement 15]. Note that this implies that the addresses used in the control plane are independent from the addresses used in the management and data planes.
15. The MPLS-TP control plane must support the physical separation of the control plane from the management and data plane, and no assumptions should be made about the state of the data plane channels from information about the control or management plane channels when they are running out-of-band [RFC5654, requirement 16].
16. A control plane must be defined to support dynamic provisioning and restoration of MPLS-TP transport paths, but its use is a network operator's choice [RFC5654, requirement 18].
17. A control plane must not be required to support the static provisioning of MPLS-TP transport paths. [RFC5654, requirement 19].
18. The MPLS-TP control plane must permit the coexistence of statically and dynamically provisioned/managed MPLS-TP transport paths within the same layer network or domain [RFC5654, requirement 20].

19. The MPLS-TP control plane should be operable in a way that is similar to the way the control plane operates in other transport-layer technologies [RFC5654, requirement 21].
20. The MPLS-TP control plane must avoid or minimize traffic impact (e.g. packet delay, reordering and loss) during network reconfiguration [RFC5654, requirement 24].
21. The MPLS-TP control plane must work across multiple homogeneous domains [RFC5654, requirement 25].
22. The MPLS-TP control plane should work across multiple non-homogeneous domains [RFC5654, requirement 26].
23. The MPLS-TP control plane must not dictate any particular physical or logical topology [RFC5654, requirement 27].
24. The MPLS-TP control plane must include support of ring topologies which may be deployed with arbitrarily interconnection, support rings of at least 16 nodes [RFC5654, requirement 27.A. and 27.B].
25. The MPLS-TP control plane must scale gracefully to support a large number of transport paths, nodes and links. That is it must be able to scale at least as well as control planes in existing transport technologies with growing and increasingly complex network topologies as well as with increasing bandwidth demands, number of customers, and number of services [RFC 5654, requirements 53 and 28].
26. The MPLS-TP control plane should not provision transport paths which contain forwarding loops [RFC5654, requirement 29].
27. The MPLS-TP control plane must support multiple client layers. (e.g. MPLS-TP, IP, MPLS, Ethernet, ATM, FR, etc.) [RFC5654, requirement 30].
28. The MPLS-TP control plane must provide a generic and extensible solution to support the transport of MPLS-TP transport paths over one or more server layer networks (such as MPLS-TP, Ethernet, SONET/SDH, OTN, etc.). Requirements for bandwidth management within a server layer network are outside the scope of this document [RFC5654, requirement 31].
29. In an environment where an MPLS-TP layer network is supporting a client layer network, and the MPLS-TP layer network is supported by a server layer network then the control plane operation of the MPLS-TP layer network must be possible without any dependencies on the server or client layer network [RFC5654, requirement 32].

30. The MPLS-TP control plane must allow for the transport of a client MPLS or MPLS-TP layer network over a server MPLS or MPLS-TP layer network [RFC5654, requirement 33].
31. The MPLS-TP control plane must allow the autonomous operation of the layers of a multi-layer network that includes an MPLS-TP layer [RFC5654, requirement 34].
32. The MPLS-TP control plane must allow the hiding of MPLS-TP layer network addressing and other information (e.g. topology) from client layer networks. However, it should be possible, at the option of the operator, to leak a limited amount of summarized information (such as SRLGs or reachability) between layers [RFC5654, requirement 35].
33. The MPLS-TP control plane must allow for the identification of a transport path on each link within and at the destination (egress) of the transport network. [RFC5654, requirement 38 and 39].
34. The MPLS-TP control plane must allow for the use of P2MP server (sub)layer capabilities as well as P2P server (sub)layer capabilities when supporting P2MP MPLS-TP transport paths [RFC5654, requirement 40].
35. The MPLS-TP control plane must be extensible in order to accommodate new types of client layer networks and services [RFC5654, requirement 41].
36. The MPLS-TP control plane should support the reserved bandwidth associated with a transport path to be increased without impacting the existing traffic on that transport path provided enough resources are available [RFC5654, requirement 42].
37. The MPLS-TP control plane should support the reserved bandwidth of a transport path to be decreased without impacting the existing traffic on that transport path, provided that the level of existing traffic is smaller than the reserved bandwidth following the decrease [RFC5654, requirement 43].
38. Requirement removed.
39. The control plane for MPLS-TP must fit within the ASON architecture. The ITU-T has defined an architecture for Automatically Switched Optical Networks (ASON) in G.8080 [ITU.G8080.2006] and G.8080 Amendment 1 [ITU.G8080.2008]. An interpretation of the ASON signaling and routing requirements in the context of GMPLS can be found in [RFC4139] and [RFC4258] [RFC5654, Section 2.4., Paragraph 2 and 3].

40. The MPLS-TP control plane must support control plane topology and data plane topology independence [RFC5654, requirement 47].
41. A failure of the MPLS-TP control plane must not interfere with the deliver of service or recovery of established transport paths [RFC5654, requirement 47].
42. The MPLS-TP control plane must be able to operate independent of any particular client or server layer control plane [RFC5654, requirement 48].
43. The MPLS-TP control plane should support, but not require, an integrated control plane encompassing MPLS-TP together with its server and client layer networks when these layer networks belong to the same administrative domain [RFC5654, requirement 49].
44. The MPLS-TP control plane must support configuration of protection functions and any associated maintenance (OAM) functions [RFC5654, requirement 50 and 7].
45. The MPLS-TP control plane must support the configuration and modification of OAM maintenance points as well as the activation/deactivation of OAM when the transport path or transport service is established or modified [RFC5654, requirement 51].
46. The MPLS-TP control plane must be capable of restarting and relearning its previous state without impacting forwarding [RFC5654, requirement 54].
47. The MPLS-TP control plane must provide a mechanism for dynamic ownership transfer of the control of MPLS-TP transport paths from the management plane to the control plane and vice versa. The number of reconfigurations required in the data plane must be minimized (preferably no data plane reconfiguration will be required) [RFC5654, requirement 55].
48. The MPLS-TP control plane must support protection and restoration mechanisms, i.e., recovery [RFC5654, requirement 52].

Note that the MPLS-TP Survivability Framework document, [TP-SURVIVE], provides additional useful information related to recovery.

49. The MPLS-TP control plane mechanisms should be identical (or as similar as possible) to those already used in existing transport networks to simplify implementation and operations. However, this must not override any other requirement [RFC5654, requirement 56 A].

50. The MPLS-TP control plane mechanisms used for P2P and P2MP recovery should be identical to simplify implementation and operation. However, this must not override any other requirement [RFC5654, requirement 56 B].
51. The MPLS-TP control plane must support recovery mechanisms that are applicable at various levels throughout the network including support for link, transport path, segment, concatenated segment and end-to-end recovery [RFC5654, requirement 57].
52. The MPLS-TP control plane must support recovery paths that meet the SLA protection objectives of the service [RFC5654, requirement 58]. Including:
  - a. Guarantee 50ms recovery times from the moment of fault detection in networks with spans less than 1200 km.
  - b. Protection of up to 100% of the traffic on the protected path.
  - c. Recovery must meet SLA requirements over multiple domains.
53. The MPLS-TP control plane should support per transport path Recovery objectives [RFC5654, requirement 59].
54. The MPLS-TP control plane must support recovery mechanisms that are applicable to any topology [RFC5654, requirement 60].
55. The MPLS-TP control plane must operate in synergy with (including coordination of timing/timer settings) the recovery mechanisms present in any client or server transport networks (for example, Ethernet, SDH, OTN, WDM) to avoid race conditions between the layers [RFC5654, requirement 61].
56. The MPLS-TP control plane must support recovery and reversion mechanisms that prevent frequent operation of recovery in the event of an intermittent defect [RFC5654, requirement 62].
57. The MPLS-TP control plane must support revertive and non-revertive protection behavior [RFC5654, requirement 64].
58. The MPLS-TP control plane must support 1+1 bidirectional protection for P2P transport paths [RFC5654, requirement 65 A].
59. The MPLS-TP control plane must support 1+1 unidirectional protection for P2P transport paths [RFC5654, requirement 65 B].

60. The MPLS-TP control plane must support 1+1 unidirectional protection for P2MP transport paths [RFC5654, requirement 65 C].
61. The MPLS-TP control plane must support the ability to share protection resources amongst a number of transport paths [RFC5654, requirement 66].
62. The MPLS-TP control plane must support 1:n bidirectional protection for P2P transport paths, and this should be the default for 1:n protection [RFC5654, requirement 67 A].
63. The MPLS-TP control plane must support 1:n unidirectional protection for P2MP transport paths [RFC5654, requirement 67 B].
64. The MPLS-TP control plane may support 1:n unidirectional protection for P2P transport paths [RFC5654, requirement 65 C].
65. The MPLS-TP control plane may support extra-traffic [RFC5654, note after requirement 67].
66. The MPLS-TP control plane should support 1:n (including 1:1) shared mesh recovery [RFC5654, requirement 68].
67. The MPLS-TP control plane must support sharing of protection resources such that protection paths that are known not to be required concurrently can share the same resources [RFC5654, requirement 69].
68. The MPLS-TP control plane must support the sharing of resources between a restoration transport path and the transport path being replaced [RFC5654, requirement 70].
69. The MPLS-TP control plane must support restoration priority so that an implementation can determine the order in which transport paths should be restored [RFC5654, requirement 71].
70. The MPLS-TP control plane must support preemption priority in order to allow restoration to displace other transport paths in the event of resource constraints [RFC5654, requirement 72 and 86].
71. The MPLS-TP control plane must support revertive and non-revertive restoration behavior [RFC5654, requirement 73].
72. The MPLS-TP control plane must support recovery being triggered by physical (lower) layer fault indications [RFC5654, requirement 74].

73. The MPLS-TP control plane must support recovery being triggered by OAM [RFC5654, requirement 75].
74. The MPLS-TP control plane must support management plane recovery triggers (e.g., forced switch, etc.) [RFC5654, requirement 76].
75. The MPLS-TP control plane must support the differentiation of administrative recovery actions from recovery actions initiated by other triggers [RFC5654, requirement 77].
76. The MPLS-TP control plane should support control plane restoration triggers (e.g., forced switch, etc.) [RFC5654, requirement 78].
77. The MPLS-TP control plane must support priority logic to negotiate and accommodate coexisting requests (i.e., multiple requests) for protection switching (e.g., administrative requests and requests due to link/node failures) [RFC5654, requirement 79].
78. The MPLS-TP control plane must support the association of protection paths and working paths (sometimes known as protection groups) [RFC5654, requirement 80].
79. The MPLS-TP control plane must support pre-calculation of recovery paths [RFC5654, requirement 81].
80. The MPLS-TP control plane must support pre-provisioning of recovery paths [RFC5654, requirement 82].
81. The MPLS-TP control plane must support the external commands defined in [RFC4427]. External controls overruled by higher priority requests (e.g., administrative requests and requests due to link/node failures) or unable to be signaled to the remote end (e.g. because of a protection state coordination fail) must be ignored/dropped [RFC5654, requirement 83].
82. The MPLS-TP control plane must permit the testing and validation of the integrity of the protection/recovery transport path [RFC5654, requirement 84 A].
83. The MPLS-TP control plane must permit the testing and validation of protection/ restoration mechanisms without triggering the actual protection/restoration [RFC5654, requirement 84 B].
84. The MPLS-TP control plane must permit the testing and validation of protection/ restoration mechanisms while the working path is in service [RFC5654, requirement 84 C].

85. The MPLS-TP control plane must permit the testing and validation of protection/ restoration mechanisms while the working path is out of service [RFC5654, requirement 84 D].
86. The MPLS-TP control plane must support the establishment and maintenance of all recovery entities and functions [RFC5654, requirement 89 A].
87. The MPLS-TP control plane must support signaling of recovery administrative control [RFC5654, requirement 89 B].
88. The MPLS-TP control plane must support protection state coordination (PSC). Since control plane network topology is independent from the data plane network topology, the PSC supported by the MPLS-TP control plane may run on resources different than the data plane resources handled within the recovery mechanism (e.g. backup) [RFC5654, requirement 89 C].
89. When present, the MPLS-TP control plane must support recovery mechanisms that are optimized for specific network topologies. These mechanisms must be interoperable with the mechanisms defined for arbitrary topology (mesh) networks to enable protection of end-to-end transport paths [RFC5654, requirement 91].
90. When present, the MPLS-TP control plane must support the control of ring topology specific recovery mechanisms [RFC5654, Section 2.5.6.1].
91. The MPLS-TP control plane must include support for differentiated services and different traffic types with traffic class separation associated with different traffic [RFC5654, requirement 110].
92. The MPLS-TP control plane must support the provisioning of services that provide guaranteed Service Level Specifications (SLS), with support for hard ([RFC3209] style) and relative ([RFC3270] style) end-to-end bandwidth guarantees [RFC5654, requirement 111].
93. The MPLS-TP control plane must support the provisioning of services which are sensitive to jitter and delay [RFC5654, requirement 112].

## 2.2. MPLS-TP Framework Derived Requirements

The following additional requirements are based on [RFC5921], [TP-P2MP-FWK] and [RFC5960]:

94. Per-packet equal cost multi-path (ECMP) load balancing is currently outside the scope of MPLS-TP [TP-DATA-PLANE , section 3.1.1., paragraph 6].
95. Penultimate hop popping (PHP) is disabled on MPLS-TP LSPs by default. [TP-DATA-PLANE , section 3.1.1., paragraph 7].
96. The MPLS-TP control plane must support both E-LSP and L-LSP MPLS DiffServ modes as specified in [RFC3270] [TP-DATA-PLANE , section 3.3.2., paragraph 12].
97. Both single-segment, see [RFC3985], and multi-segment PWs, see [RFC5659], shall be supported by the MPLS-TP control plane. MPLS-TP shall use the definition of multi-segment PWs as defined by the IETF [RFC5921, section 3.4.4].
98. The MPLS-TP control plane must support the control of PWs and their associated labels [RFC5921, section 3.4.4].
99. The MPLS-TP control plane must support network layer clients, i.e., clients whose traffic is transported over an MPLS-TP network without the use of PWs [RFC5921, section 3.4.5].
  - a. The MPLS-TP control plane must support the use of network layer protocol-specific LSPs and labels. [RFC5921, section 3.4.5.]
  - b. The MPLS-TP control plane must support the use of a client service-specific LSPs and labels. [RFC5921, section 3.4.5.]
100. The MPLS-TP control plane is based on the GMPLS control plane for MPLS-TP LSPs. More specifically, GMPLS RSVP-TE [RFC3473] and related extensions are used for LSP signaling, and GMPLS OSPF-TE [RFC5392] and ISIS-TE [RFC5316] are used for routing [RFC5921, section 3.9].
101. The MPLS-TP control plane is based on the MPLS control plane for PWs, and more specifically, targeted LDP (T-LDP) [RFC4447] is used for PW signaling [RFC5921, section 3.9., paragraph 5].
102. The MPLS-TP control plane must ensure its own survivability and to enable it to recover gracefully from failures and degradations. These include graceful restart and hot redundant configurations [RFC5921, section 3.9., paragraph 16].

- 103. The MPLS-TP control plane must support linear, ring and meshed protection schemes [RFC5921, section 3.12., paragraph 3].
- 104. The MPLS-TP control plane must support the control of SPMEs (hierarchical LSPs) for new or existing end-to-end LSPs [RFC5921, section 3.12., paragraph 7].

### 2.3. OAM Framework Derived Requirements

The following additional requirements are based on [RFC5860] and [TP-OAM]:

- 105. The MPLS-TP control plane must support the capability to enable/disable OAM functions as part of service establishment [RFC5860, section 2.1.6., paragraph 1]. Note that OAM functions are applicable regardless of the label stack depth (i.e., level of LSP hierarchy or PW) [RFC5860, section 2.1.1., paragraph 3].
- 106. The MPLS-TP control plane must support the capability to enable/disable OAM functions after service establishment. In such cases, the customer must not perceive service degradation as a result of OAM enabling/disabling [RFC5860, section 2.1.6., paragraph 1 and 2].
- 107. The MPLS-TP control plane must support dynamic control of any of the existing IP/MPLS and PW OAM protocols (e.g., LSP-Ping [RFC4379], MPLS-BFD [RFC5884], VCCV [RFC5085], and VCCV-BFD [RFC5885]) [RFC5860, section 2.1.4., paragraph 2].
- 108. The MPLS-TP control plane must allow for the ability to support experimental OAM functions. These functions must be disabled by default [RFC5860, section 2.2., paragraph 2].
- 109. The MPLS-TP control plane must support the choice of which (if any) OAM function(s) to use and to which PW, LSP or Section it applies [RFC5860, section 2.2., paragraph 3].
- 110. The MPLS-TP control plane must allow (e.g., enable/disable) mechanisms that support the localization of faults and the notification of appropriate nodes. [RFC5860, section 2.2.1., paragraph 1].
- 111. The MPLS-TP control plane may support mechanisms that permit the service provider to be informed of a fault or defect affecting the service(s) it provides, even if the fault or defect is located outside of his domain [RFC5860, section 2.2.1., paragraph 2].

112. Information exchange between various nodes involved in the MPLS-TP control plane should be reliable such that, for example, defects or faults are properly detected or that state changes are effectively known by the appropriate nodes [RFC5860, section 2.2.1., paragraph 3].
113. The MPLS-TP control plane must provide functionality to control an End Point's ability to monitor the liveness of a PW, LSP, or Section [RFC5860, section 2.2.2., paragraph 1].
114. The MPLS-TP control plane must provide functionality to control an End Point's ability to determine whether or not it is connected to specific End Point(s) by means of the expected PW, LSP, or Section [RFC5860, section 2.2.3., paragraph 1].
  - a. The MPLS-TP control plane must provide mechanisms to control an End Point's ability to perform this function proactively [RFC5860, section 2.2.3., paragraph 2].
  - b. The MPLS-TP control plane must provide mechanisms to control an End Point's ability to perform this function on-demand [RFC5860, section 2.2.3., paragraph 3].
115. The MPLS-TP control plane must provide functionality to control diagnostic testing on a PW, LSP or Section [RFC5860, section 2.2.5., paragraph 1].
  - a. The MPLS-TP control plane must provide mechanisms to control the performance of this function on-demand [RFC5860, section 2.2.5., paragraph 2].
116. The MPLS-TP control plane must provide functionality to enable an End Point to discover the Intermediate (if any) and End Point(s) along a PW, LSP or Section, and more generally to trace (record) the route of a PW, LSP or Section [RFC5860, section 2.2.4., paragraph 1].
  - a. The MPLS-TP control plane must provide mechanisms to control the performance of this function on-demand [RFC5860, section 2.2.4., paragraph 2].
117. The MPLS-TP control plane must provide functionality to enable an End Point of a PW, LSP or Section to instruct its associated End Point(s) to lock the PW, LSP or Section [RFC5860, section 2.2.6., paragraph 1].
  - a. The MPLS-TP control plane must provide mechanisms to control the performance of this function on-demand [RFC5860, section 2.2.6., paragraph 2].

118. The MPLS-TP control plane must provide functionality to enable an Intermediate Point of a PW or LSP to report, to an End Point of that same PW or LSP, a lock condition indirectly affecting that PW or LSP [RFC5860, section 2.2.7., paragraph 1].
  - a. The MPLS-TP control plane must provide mechanisms to control the performance of this function proactively [RFC5860, section 2.2.7., paragraph 2].
119. The MPLS-TP control plane must provide functionality to enable an Intermediate Point of a PW or LSP to report, to an End Point of that same PW or LSP, a fault or defect condition affecting that PW or LSP [RFC5860, section 2.2.8., paragraph 1].
  - a. The MPLS-TP control plane must provide mechanisms to control the performance of this function proactively [RFC5860, section 2.2.8., paragraph 2].
120. The MPLS-TP control plane must provide functionality to enable an End Point to report, to its associated End Point, a fault or defect condition that it detects on a PW, LSP or Section for which they are the End Points [RFC5860, section 2.2.9., paragraph 1].
  - a. The MPLS-TP control plane must provide mechanisms to control the performance of this function proactively [RFC5860, section 2.2.9., paragraph 2].
121. The MPLS-TP control plane must provide functionality to enable the propagation, across an MPLS-TP network, of information pertaining to a client defect or fault condition detected at an End Point of a PW or LSP, if the client layer mechanisms do not provide an alarm notification/propagation mechanism [RFC5860, section 2.2.10., paragraph 1].
  - a. The MPLS-TP control plane must provide mechanisms to control the performance of this function proactively [RFC5860, section 2.2.10., paragraph 2].
122. The MPLS-TP control plane must provide functionality to enable the control of quantification of packet loss ratio over a PW, LSP or Section [RFC5860, section 2.2.11., paragraph 1].
  - a. The MPLS-TP control plane must provide mechanisms to control the performance of this function proactively and on-demand [RFC5860, section 2.2.11., paragraph 4].
123. The MPLS-TP control plane must provide functionality to control the quantification and reporting of the one-way, and if appropriate, the two-way, delay of a PW, LSP or Section [RFC5860, section 2.2.12., paragraph 1].

- a. The MPLS-TP control plane must provide mechanisms to control the performance of this function proactively and on-demand [RFC5860, section 2.2.12., paragraph 6].
124. The MPLS-TP control plane must support the configuration of OAM functional components which include MEs and MEGs as instantiated in MEPs, MIPs and SPMEs [TP-OAM, section 3.6].
125. For dynamically established transport paths, the control plane must support the configuration of OAM operations [TP-OAM, section 5].
- a. The MPLS-TP control plane must provide mechanisms to configure proactive monitoring for a MEG at, or after, transport path creation time.
  - b. The MPLS-TP control plane must provide mechanisms to configure the operational characteristics of in-band measurement transactions (e.g., CV, LM etc.) are configured at the MEPs (associated with a transport path).
  - c. The MPLS-TP control plane may provide mechanisms to configure server layer event reporting by intermediate nodes.
  - d. The MPLS-TP control plane may provide mechanisms to configure the reporting of measurements resulting from proactive monitoring.
126. The MPLS-TP control plane must support the control of the loss of continuity (LOC) traffic block consequent action [TP-OAM, section 5.1.2., paragraph 4].
127. For dynamically established transport paths that have a proactive CC-V function enabled, the control plane must support the signaling of the following MEP configuration information [TP-OAM, section 5.1.3]:
- a. The MPLS-TP control plane must provide mechanisms to configure the MEG identifier to which the MEP belongs.
  - b. The MPLS-TP control plane must provide mechanisms to configure a MEP's own identity inside a MEG.
  - c. The MPLS-TP control plane must provide mechanisms to configure the list of the other MEPs in the MEG.

- d. The MPLS-TP control plane must provide mechanisms to configure the CC-V transmission rate / reception period (covering all application types).
- 128. The MPLS-TP control plane must provide mechanisms to configure the generation of Alarm Indication Signal (AIS) packets for each MEG [TP-OAM, section 5.3., paragraph 9].
  - 129. The MPLS-TP control plane must provide mechanisms to configure the generation of Locked Report (LKR) packets for each MEG [TP-OAM, section 5.4., paragraph 9].
  - 130. The MPLS-TP control plane must provide mechanisms to configure the use of proactive Packet Loss Measurement (LM), and the transmission rate and PHB class associated with the LM OAM packets originating from a MEP [TP-OAM, section 5.5.1., paragraph 1].
  - 131. The MPLS-TP control plane must provide mechanisms to configure the use of proactive Packet Delay Measurement (DM), and the transmission rate and PHB class associated with the DM OAM packets originating from a MEP [TP-OAM, section 5.6.1., paragraph 1].
  - 132. The MPLS-TP control plane must provide mechanisms to configure the use of Client Failure Indication (CFI), and the transmission rate and PHB class associated with the CFI OAM packets originating from a MEP [TP-OAM, section 5.7.1., paragraph 1].
  - 133. The MPLS-TP control plane should provide mechanisms to control the use of on-demand CV packets [TP-OAM, section 6.1].
    - a. The MPLS-TP control plane should provide mechanisms to configure the number of packets to be transmitted/received in each burst of on-demand CV packets and their packet size [TP-OAM, section 6.1.1, paragraph 1].
    - b. When an on-demand CV packet is used to check connectivity toward a target MIP, the MPLS-TP control plane should provide mechanisms to configure the number of hops to reach the target MIP [TP-OAM, section 6.1.1, paragraph 2].
    - c. The MPLS-TP control plane should provide mechanisms to configure the PHB of on-demand CV packets [TP-OAM, section 6.1.1, paragraph 3].

134. The MPLS-TP control plane should provide mechanisms to control the use of on-demand LM, including configuration of the beginning and duration of the LM procedures, the transmission rate and PHB associated with the LM OAM packets originating from a MEP. [TP-OAM, section 6.2.1.]
135. The MPLS-TP control plane should provide mechanisms to control the use of Throughput estimation [TP-OAM, section 6.3.1].
136. The MPLS-TP control plane should provide mechanisms to control the use of on-demand DM, including configuration of the beginning and duration of the DM procedures, the transmission rate and PHB associated with the DM OAM packets originating from a MEP. [TP-OAM, section 6.5.1.]

#### 2.4. Security Requirements

There are no specific MPLS-TP control plane security requirements. The existing framework for MPLS and GMPLS security is documented in [RFC5920] and that document applies equally to MPLS-TP.

#### 2.5. Identifier Requirements

The following are requirements based on [TP-IDENTIFIERS]:

137. The MPLS-TP control plane must support MPLS-TP point to point tunnel identifiers of the forms defined in [TP-IDENTIFIERS, Section 5.1].
138. The MPLS-TP control plane must support MPLS-TP LSP identifiers of the forms defined in [TP-IDENTIFIERS, Section 5.2], and the mappings to GMPLS as defined in [TP-IDENTIFIERS, Section 5.3].
139. The MPLS-TP control plane must support Pseudowire path identifiers of the form defined in [TP-IDENTIFIERS, Section 6].
140. The MPLS-TP control plane must support MEG\_IDs for LSPs and PWs as defined in [TP-IDENTIFIERS, Section 7.1.1].
141. The MPLS-TP control plane must support IP compatible MEG\_IDs for LSPs and PWs as defined [TP-IDENTIFIERS, Section 7.1.2].
142. The MPLS-TP control plane must support MEP\_IDs for LSPs and PWs of the forms defined in [TP-IDENTIFIERS, Section 7.2.1].
143. The MPLS-TP control plane must support IP based MEP\_IDs for MPLS-TP LSP of the forms defined in [TP-IDENTIFIERS, Section 7.2.2.1].

144. The MPLS-TP control plane must support IP based MEP\_IDs for Pseudowires of the form defined in [TP-IDENTIFIERS, Section 7.2.2.2].

### 3. Relationship of PWs and TE LSPs

The data plane relationship between PWs and LSPs is inherited from standard MPLS and is reviewed in the MPLS-TP Framework [RFC5921]. Likewise, the control plane relationship between PWs and LSPs is inherited from standard MPLS. This relationship is reviewed in this document. The relationship between the PW and LSP control planes in MPLS-TP is the same as the relationship found in the PWE3 Maintenance Reference Model as presented in the PWE3 Architecture, see Figure 6 of [RFC3985]. The PWE3 Architecture [RFC3985] states: "the PWE3 protocol-layering model is intended to minimize the differences between PWs operating over different PSN types." Additionally, PW control (maintenance) takes place separately from LSP signaling. [RFC4447] and [MS-PW-DYNAMIC] provide such extensions for the use of LDP as the control plane for PWs. This control can provide PW control without providing LSP control.

In the context of MPLS-TP, LSP tunnel signaling is provided via GMPLS RSVP-TE. While RSVP-TE could be extended to support PW control much as LDP was extended in [RFC4447], such extensions are out of scope of this document. This means that the control of PWs and LSPs will operate largely independently. The main coordination between LSP and PW control will occur within the nodes that terminate PWs, or PW segments. See Section 5.3.2 for an additional discussion on such coordination.

It is worth noting that the control planes for PWs and LSPs may be used independently, and that one may be employed without the other. This translates into the four possible scenarios: (1) no control plane is employed; (2) a control plane is used for both LSPs and PWs; (3) a control plane is used for LSPs, but not PWs; (4) a control plane is used for PWs, but not LSPs.

The PW and LSP control planes, collectively, must satisfy the MPLS-TP control plane requirements reviewed in this document. When client services are provided directly via LSPs, all requirements must be satisfied by the LSP control plane. When client services are provided via PWs, the PW and LSP control planes can operate in combination and some functions may be satisfied via the PW control plane while others are provided to PWs by the LSP control plane. For example, to support the recovery functions described in [TP-SURVIVE] this document focuses on the control of the recovery functions at the LSP layer. PW based recovery is under development at this time and may be used once defined.

#### 4. TE LSPs

MPLS-TP uses Generalized MPLS (GMPLS) signaling and routing, see [RFC3945], as the control plane for LSPs. The GMPLS control plane is based on the MPLS control plane. GMPLS includes support for MPLS labeled data and transport data planes. GMPLS includes most of the transport centric features required to support MPLS-TP LSPs. This section will first review the features of GMPLS relevant to MPLS-TP LSPs, then identify how specific requirements can be met using existing GMPLS functions, and will conclude with extensions that are anticipated to support the remaining MPLS-TP control plane requirements.

##### 4.1. GMPLS Functions and MPLS-TP LSPs

This section reviews how existing GMPLS functions can be applied to MPLS-TP.

###### 4.1.1. In-Band and Out-Of-Band Control

GMPLS supports both in-band and out-of-band control. The terms in-band and out-of-band, in the context of this document, refer to the relationship of the control plane relative to the management and data planes. The terms may be used to refer to the control plane independent of the management plane, or to both of them in concert. The remainder of this section describes the relationship of the control plane to the management and data planes.

There are multiple uses of both terms in-band and out-of-band. The terms may relate to a channel, a path or a network. Each of these can be used independently or in combination. Briefly, some typical usage of the terms are as follows:

- o In-band

This term is used to refer to cases where control plane traffic is sent in the same communication channel used to transport associated user data or management traffic. IP, MPLS, and Ethernet networks are all examples where control traffic is typically sent in-band with the data traffic. An example of this case in the context of MPLS-TP is where control plane traffic is sent via the MPLS Generic Associated Channel (G-ACh), see [RFC5586], using the same LSP as controlled user traffic.

- o Out-of-band, in-fiber

This term is used to refer to cases where control plane traffic is sent using a different communication channel from the associated data or management traffic, and the control communication channel resides in the same fiber as either the management or data traffic. An example of this case in the

context of MPLS-TP is where control plane traffic is sent via the G-ACh using a dedicated LSP on the same link (interface) which carries controlled user traffic.

- o Out-of-band, aligned topology

This term is used to refer to the cases where control plane traffic is sent using a different communication channel from the associated data or management traffic, and the control traffic follows the same node-to-node path as either the data or management traffic.

Such topologies are usually supported using a parallel fiber or other configurations where multiple data channels are available and one is (dynamically) selected as the control channel. An example of this case in the context of MPLS-TP is where control plane traffic is sent along the same node pairs, but not necessarily the same links (interfaces), as the corresponding controlled user traffic.

- o Out-of-band, independent topology

This term is used to refer to the cases where control plane traffic is sent using a different communication channel from the associated data or management traffic, and the control traffic may follow a path that is completely independent of the data traffic.

Such configurations are a superset of the other cases and do not preclude the use of in-fiber or aligned topology links, but alignment is not required. An example of this case in the context of MPLS-TP is where control plane traffic is sent between controlling nodes using any available path and links, completely without regard for the path(s) taken by corresponding management or user traffic.

In the context of MPLS-TP requirements, requirement 14 (see Section 2 above) can be met using out-of-band in-fiber or aligned topology types of control. Requirement 15 can only be met by using Out-of-band, independent topology. Some expect the G-ACh to be used extensively in MPLS-TP networks to support the MPLS-TP control (and management) planes.

#### 4.1.2. Addressing

MPLS-TP reuses and supports the addressing mechanisms supported by MPLS. The MPLS-TP Identifiers document, see [TP-IDENTIFIERS], provides additional context on how IP addresses are used within MPLS-TP. MPLS, and consequently, MPLS-TP uses the IPv4 and IPv6 address families to identify MPLS-TP nodes by default for network management and signaling purposes. The address spaces and neighbor adjacencies in the control, management and data planes used in an MPLS-TP network

may be completely separated or combined at the discretion of an MPLS-TP operator and based on the equipment capabilities of a vendor. The separation of the control and management planes from the data plane allows each plane to be independently addressable. Each plane may use addresses that are not mutually reachable, e.g., it is likely that the data plane will not be able to reach an address from the management or control planes and vice versa. Each plane may also use a different address family. It is even possible to reuse addresses in each plane, but this is not recommended as it may lead to operational confusion. As previously mentioned, the G-ACh mechanism defined in [RFC5586] is expected to be used extensively in MPLS-TP networks to support the MPLS-TP control (and management) planes.

#### 4.1.3. Routing

Routing support for MPLS-TP LSPs is based on GMPLS routing. GMPLS routing builds on TE routing and has been extended to support multiple switching technologies per [RFC3945] and [RFC4202] as well as multiple levels of packet switching (PSC) within a single network. IS-IS extensions for GMPLS are defined in [RFC5307] and [RFC5316], which build on the TE extensions to IS-IS defined in [RFC5305]. OSPF extensions for GMPLS are defined in [RFC4203] and [RFC5392], which build on the TE extensions to OSPF defined in [RFC3630]. The listed RFCs should be viewed as a starting point rather than an comprehensive list as there are other IS-IS and OSPF extensions, as defined in IETF RFCs, that can be used within an MPLS-TP network.

#### 4.1.4. TE LSPs and Constraint-Based Path Computation

Both MPLS and GMPLS allow for traffic engineering and constraint-based path computation. MPLS path computation provides paths for MPLS-TE unidirectional P2P and P2MP LSPs. GMPLS path computation adds bidirectional LSPs, explicit recovery path computation as well as support for the other functions discussed in this section.

Both MPLS and GMPLS path computation allow for the restriction of path selection based on the use of Explicit Route Objects (EROs) and other LSP attributes, see [RFC3209] and [RFC3473]. In all cases, no specific algorithm is standardized by the IETF. This is anticipated to continue to be the case for MPLS-TP LSPs.

##### 4.1.4.1. Relation to PCE

Path Computation Element (PCE) Based approaches, see [RFC4655], may be used for path computation of a GMPLS LSP, and consequently an MPLS-TP LSP, across domains and in a single domain. In cases where PCE is used, the PCE Communication Protocol (PCEP), see [RFC5440], will be used to communicate PCE requests and responses. MPLS-TP

specific extensions to PCEP are currently out of scope of the MPLS-TP project and this document.

#### 4.1.5. Signaling

GMPLS signaling is defined in [RFC3471] and [RFC3473], and is based on RSVP-TE [RFC3209]. CR-LDP based GMPLS, [RFC3472] is no longer under active development within the IETF, i.e., it is deprecated, and must not be used for MPLS-TP. In general, all RSVP-TE extensions that apply to MPLS may also be used for GMPLS and consequently MPLS-TP. Most notably this includes support for P2MP signaling as defined in [RFC4875].

GMPLS signaling includes a number of MPLS-TP required functions. Notably support for out-of-band control, bidirectional LSPs, and independent control and data plane fault management. There are also numerous other GMPLS and MPLS extensions that can be used to provide specific functions in MPLS-TP networks. Specific references are provided below.

#### 4.1.6. Unnumbered Links

Support for unnumbered links (i.e., links that do not have IP addresses) is permitted in MPLS-TP and its usage is at the discretion of the network operator. Support for unnumbered links is included for routing in [RFC4203] for OSPF and [RFC5307] for IS-IS, and for signaling in [RFC3477].

#### 4.1.7. Link Bundling

Link bundling provides a local construct that can be used to improve scaling of TE routing when multiple data links are shared between node pairs. Link bundling for MPLS and GMPLS networks is defined in [RFC4201]. Link bundling may be used in MPLS-TP networks and its use is at the discretion of the network operator.

#### 4.1.8. Hierarchical LSPs

This section reuses text from [HIERARCHY-BIS].

[RFC3031] describes how MPLS labels may be stacked so that LSPs may be nested with one LSP running through another. This concept of Hierarchical LSPs (H-LSPs) is formalized in [RFC4206] with a set of protocol mechanisms for the establishment of a hierarchical LSP that can carry one or more other LSPs.

[RFC4206] goes on to explain that a hierarchical LSP may carry other

LSPs only according to their switching types. This is a function of the way labels are carried. In a packet switch capable (PSC) network, the hierarchical LSP can carry other PSC LSPs using the MPLS label stack.

Signaling mechanisms defined in [RFC4206] allow a hierarchical LSP to be treated as a single hop in the path of another LSP. This mechanism is also sometimes known as "non-adjacent signaling", see [RFC4208].

A Forwarding Adjacency (FA) is defined in [RFC4206] as a data link created from an LSP and advertised in the same instance of the control plane that advertises the TE links from which the LSP is constructed. The LSP itself is called an FA-LSP. FA LSPs are analogous to MPLS-TP Sections as discussed in [RFC5960].

Thus, a hierarchical LSP may form an FA such that it is advertised as a TE link in the same instance of the routing protocol as was used to advertise the TE links that the LSP traverses.

As observed in [RFC4206] the nodes at the ends of an FA would not usually have a routing adjacency.

LSP hierarchy is expected to play an important role in MPLS-TP networks, particularly in the context of scaling and recovery as well as supporting SPMEs.

#### 4.1.9. LSP Recovery

GMPLS defines RSVP-TE extensions in support for end-to-end GMPLS LSPs recovery in [RFC4872], and segment recovery in [RFC4873]. GMPLS segment recovery provides a superset of the function in end-to-end recovery. End-to-end recovery can be viewed as a special case of segment recovery where there is a single recovery domain whose borders coincide with the ingress and egress of the LSP, although specific procedures are defined.

The five defined types of recovery defined in GMPLS are:

- 1+1 bidirectional protection for P2P LSPs
- 1+1 unidirectional protection for P2MP LSPs
- 1:n (including 1:1) protection with or without extra traffic
- Rerouting without extra traffic (sometimes known as soft rerouting), including shared mesh restoration
- Full LSP rerouting

Recovery for MPLS-TP LSPs as discussed in [TP-SURVIVE], is signaled using the mechanism defined in [RFC4872] and [RFC4873]. Note that when MEPs are required for the OAM CC function and the MEPs exist at LSP transit nodes, each MEP is instantiated at a hierarchical LSP end point, and protection is provided end-to-end for the hierarchical

LSP. (Protection can be signaled using either [RFC4872] or [RFC4873] defined procedures.) The use of Notify messages to trigger protection switching and recovery is not required in MPLS-TP as this function is expected to be supported via OAM. However, its use is not precluded.

#### 4.1.10. Control Plane Reference Points (E-NNI, I-NNI, UNI)

The majority of GMPLS control plane related RFCs define the control plane from the context of an internal network-to-network interface (I-NNI). In the MPLS-TP context, some operators may choose to deploy signaled interfaces across user-to-network (UNI) interfaces and across inter-provider, external network-to-network (E-NNI), interfaces. Such support is embodied in [RFC4208] for UNIs and [RFC5787] for routing areas in support of E-NNIs. This work may require extensions in order to meet the specific needs of an MPLS-TP UNI and E-NNI.

#### 4.2. OAM, MEP (Hierarchy), MIP Configuration and Control

MPLS-TP is being defined to support a comprehensive set of MPLS-TP OAM functions. The MPLS-TP control plane will not itself provide OAM functions, but it will be used to instantiate and otherwise control MPLS-TP OAM functions.

Specific OAM requirements for MPLS-TP are documented in [RFC5860]. This document also states that it is also required that the control plane be able to configure and control OAM entities. This requirement is not yet addressed by the existing RFCs, but such work is now underway, e.g., [CCAMP-OAM-FWK] and [CCAMP-OAM-EXT].

Many OAM functions occur on a per-LSP basis, are typically in-band, and are initiated immediately after LSP establishment. Hence, it is desirable that such functions be established and activated via the same control plane signaling used to set up the LSP, as this effectively synchronizes OAM with the LSP lifetime and avoids the extra overhead and potential errors associated with separate OAM configuration mechanisms.

##### 4.2.1. Management Plane Support

There is no MPLS-TP requirement for a standardized management interface to the MPLS-TP control plane. That said, MPLS and GMPLS support a number of standardized management functions. These include the MPLS-TE/GMPLS TE Database Management Information Base (MIB), [TE-MIB]; the MPLS-TE MIB, [RFC3812]; the MPLS LSR MIB, [RFC3813]; the GMPLS TE MIB [RFC4802]; and the GMPLS LSR MIB, [RFC4803]. These MIBs may be used in MPLS-TP networks.

#### 4.2.1.1. Recovery Triggers

The GMPLS control plane allows for management plane recovery triggers and directly supports control plane recovery triggers. Support for control plane recovery triggers is defined in [RFC4872] which refers to the triggers as "Recovery Commands". These commands can be used with both end-to-end and segment recovery, but are always controlled on an end-to-end basis. The recovery triggers/commands defined in [RFC4872] are:

- a. Lockout of recovery LSP
- b. Lockout of normal traffic
- c. Forced switch for normal traffic
- d. Requested switch for normal traffic
- e. Requested switch for recovery LSP

Note that control plane triggers are typically invoked in response to a management plane request at the ingress.

#### 4.2.1.2. Management Plane / Control Plane Ownership Transfer

In networks where both control plane and management plane are provided, LSP provisioning can be done either by control plane or management plane. As mentioned in the requirements section above, it must be possible to transfer, or handover, a management plane created LSP to the control plane domain and vice versa. [RFC5493] defines the specific requirements for an LSP ownership handover procedure. It must be possible for the control plane to provide the management plane, in a reliable manner, with the status or result of an operation performed by the management plane. This notification may be either synchronous or asynchronous with respect to the operation. Moreover, it must be possible for the management plane to monitor the status of the control plane, for example the status of a TE Link, its available resources, etc. This monitoring may be based on queries initiated by the management plane or on notifications generated by the control plane. A mechanism must be made available by the control plane to the management plane to log control plane LSP related operation, that is, it must be possible from the NMS to have a clear view of the life, (traffic hit, action performed, signaling etc.) of a given LSP. The LSP handover procedure for MPLS-TP LSPs is supported via [RFC5852].

#### 4.3. GMPLS and MPLS-TP Requirements Table

The following table shows how the MPLS-TP control plane requirements can be met using the existing GMPLS control plane (which builds on the MPLS control plane). Areas where additional specifications are required are also identified. The table lists references based on the control plane requirements as identified and numbered above in section 2.

Req #	References
1	Generic requirement met by using Standards Track RFCs
2	[RFC3945], [RFC4202], [RFC3473], [RFC4203], [RFC5307]
3	[RFC5145] + Formal Definition (See Section 4.4.1)
4	Generic requirement met by using Standards Track RFCs
5	[RFC3945], [RFC4202], [RFC3473], [RFC4203], [RFC5307]
6	[RFC3471], [RFC3473], [RFC4875]
7	[RFC3471], [RFC3473] + Associated bidirectional LSPs (See Section 4.4.2)
8	[RFC4875]
9	[RFC3473]
10	Associated bidirectional LSPs (See Section 4.4.2)
11	Associated bidirectional LSPs (See Section 4.4.2)
12	[RFC3473]
13	[RFC5467] (Currently Experimental, See Section 4.4.3)
14	[RFC3945], [RFC3473], [RFC4202], [RFC4203], [RFC5307]
15	[RFC3945], [RFC3473], [RFC4202], [RFC4203], [RFC5307]
16	[RFC3945], [RFC4202], [RFC3473], [RFC4203], [RFC5307]
17	[RFC3945], [RFC4202] + proper vendor implementation
18	[RFC3945], [RFC4202] + proper vendor implementation
19	[RFC3945], [RFC4202]
20	[RFC3473]
21	[RFC3945], [RFC4202], [RFC3473], [RFC4203], [RFC5307], [RFC5151]
22	[RFC3945], [RFC4202], [RFC3473], [RFC4203], [RFC5307], [RFC5151]
23	[RFC3945], [RFC4202], [RFC3473], [RFC4203], [RFC5307]
24	[RFC3945], [RFC4202], [RFC3473], [RFC4203], [RFC5307]
25	[RFC3945], [RFC4202], [RFC3473], [RFC4203], [RFC5307], [HIERARCHY-BIS]
26	[RFC3473], [RFC4875]
27	[RFC3473], [RFC4875]
28	[RFC3945], [RFC3471], [RFC4202]
29	[RFC3945], [RFC4202], [RFC3473], [RFC4203], [RFC5307]
30	[RFC3945], [RFC3471], [RFC4202]
31	[RFC3945], [RFC3471], [RFC4202]
32	[RFC4208], [RFC4974], [RFC5787], [RFC6001]
33	[RFC3473], [RFC4875]
34	[RFC4875]
35	[RFC3945], [RFC4202], [RFC3473], [RFC4203], [RFC5307]
36	[RFC3473], [RFC3209] (Make-before-break)
37	[RFC3473], [RFC3209] (Make-before-break)
38	
39	[RFC4139], [RFC4258], [RFC5787]
40	[RFC3945], [RFC4202], [RFC3473], [RFC4203], [RFC5307]
41	[RFC3473], [RFC5063]
42	[RFC3945], [RFC3471], [RFC4202], [RFC4208]
43	[RFC3945], [RFC3471], [RFC4202]
44	[RFC4872], [RFC4873], [CCAMP-OAM-FWK], [CCAMP-OAM-EXT]

45	[HIERARCHY-BIS], [CCAMP-OAM-FWK], [CCAMP-OAM-EXT]
46	[RFC3473], [RFC4203], [RFC5307], [RFC5063]
47	[RFC5493]
48	[RFC4872], [RFC4873]
49	[RFC3945], [RFC3471], [RFC4202]
50	[RFC4872], [RFC4873] + Recovery for P2MP (see Sec. 4.4.4)
51	[RFC4872], [RFC4873]
52	[RFC4872], [RFC4873] + proper vendor implementation
53	[RFC4872], [RFC4873], [GMPLS-PS]
54	[RFC4872], [RFC4873]
55	[RFC3473], [RFC4872], [RFC4873], [GMPLS-PS]
	Timers are a local implementation matter
56	[RFC4872], [RFC4873], [GMPLS-PS] +
	implementation of timers
57	[RFC4872], [RFC4873], [GMPLS-PS]
58	[RFC4872], [RFC4873]
59	[RFC4872], [RFC4873]
60	[RFC4872], [RFC4873]
61	[RFC4872], [RFC4873], [HIERARCHY-BIS]
62	[RFC4872], [RFC4873]
63	[RFC4872], [RFC4873] + Recovery for P2MP (see Sec. 4.4.4)
64	[RFC4872], [RFC4873]
65	[RFC4872], [RFC4873]
66	[RFC4872], [RFC4873]
67	[RFC4872], [RFC4873], [HIERARCHY-BIS]
68	[RFC4872], [RFC4873]
69	[RFC3473], [RFC4872], [RFC4873]
70	[RFC3473]
71	[RFC3473], [RFC4872], [GMPLS-PS]
72	[RFC3473], [RFC4872]
73	[RFC4872], [RFC4873], [CCAMP-OAM-FWK], [CCAMP-OAM-EXT]
74	[RFC4426], [RFC4872], [RFC4873]
75	[RFC4426], [RFC4872], [RFC4873]
76	[RFC4426], [RFC4872], [RFC4873]
77	[RFC4426], [RFC4872], [RFC4873]
78	[RFC4426], [RFC4872], [RFC4873]
79	[RFC4426], [RFC4872], [RFC4873] + vendor implementation
80	[RFC4426], [RFC4872], [RFC4873]
81	[RFC4426], [RFC4872], [RFC4873]
82	[RFC4872], [RFC4873] + Testing control (See Sec. 4.4.5)
83	[RFC4872], [RFC4873] + Testing control (See Sec. 4.4.5)
84	[RFC4872], [RFC4873] + Testing control (See Sec. 4.4.5)
85	[RFC4872], [RFC4873] + Testing control (See Sec. 4.4.5)
86	[RFC4872], [RFC4873], [CCAMP-OAM-FWK], [CCAMP-OAM-EXT]
87	[RFC4872], [RFC4873]
88	[RFC4872], [RFC4873]
89	[RFC4872], [RFC4873], [TP-RING]
90	[RFC4872], [RFC4873], [TP-RING]
91	[RFC3270], [RFC3473], [RFC4124] + GMPLS Usage (See 4.4.6)
92	[RFC3945], [RFC4202], [RFC3473], [RFC4203], [RFC5307]
93	[RFC3945], [RFC3473], [RFC2210], [RFC2211], [RFC2212]

94	Generic requirement on data plane (correct implementation)
95	[RFC3473], [NO-PHP]
96	[RFC3270], [RFC3473], [RFC4124] + GMPLS Usage (See 4.4.6)
97	PW only requirement, see PW Requirements Table (5.2)
98	PW only requirement, see PW Requirements Table (5.2)
99	[RFC3945], [RFC3473], [HIERARCHY-BIS]
100	[RFC3945], [RFC4202], [RFC3473], [RFC4203], [RFC5307] + [RFC5392] and [RFC5316]
101	PW only requirement, see PW Requirements Table (5.2)
102	[RFC3473], [RFC4203], [RFC5307], [RFC5063]
103	[RFC4872], [RFC4873], [TP-RING]
104	[RFC3945], [RFC3473], [HIERARCHY-BIS]
105	[CCAMP-OAM-FWK], [CCAMP-OAM-EXT]
106	[RFC3473], [CCAMP-OAM-FWK], [CCAMP-OAM-EXT]
107	[CCAMP-OAM-FWK], [CCAMP-OAM-EXT]
108	[CCAMP-OAM-FWK], [CCAMP-OAM-EXT] + (See Sec. 4.4.5)
109	[CCAMP-OAM-FWK], [CCAMP-OAM-EXT]
110	[RFC3473], [RFC4872], [RFC4873]
111	[RFC3473], [RFC4872], [RFC4873]
112	[RFC3473], [RFC4783]
113	[CCAMP-OAM-FWK], [CCAMP-OAM-EXT]
114	[CCAMP-OAM-FWK], [CCAMP-OAM-EXT] + (See Sec. 4.4.5)
115	[CCAMP-OAM-FWK], [CCAMP-OAM-EXT] + (See Sec. 4.4.5)
116	[RFC3473]
117	[RFC4426], [RFC4872], [RFC4873]
118	[RFC3473], [RFC4872], [RFC4873]
119	[RFC3473], [RFC4783]
120	[RFC3473]
121	[RFC3473], [RFC4783]
122	[CCAMP-OAM-FWK], [CCAMP-OAM-EXT] + (See Sec. 4.4.5)
123	[CCAMP-OAM-FWK], [CCAMP-OAM-EXT] + (See Sec. 4.4.5)
124	[CCAMP-OAM-FWK], [CCAMP-OAM-EXT], [HIERARCHY-BIS]
125 -	
136	[CCAMP-OAM-FWK], [CCAMP-OAM-EXT] + (See Sec. 4.4.5)
137a	[RFC3473]
137b	[RFC3473] + (See Sec. 4.4.7)
138a	[RFC3473]
138b	[RFC3473] + (See Sec. 4.4.7)
139	PW only requirement, see PW Requirements Table (5.2)
140 -	
144	[CCAMP-OAM-FWK], [CCAMP-OAM-EXT] + (See Sec. 4.4.8)

+=====+

#### 4.4. Anticipated MPLS-TP Related Extensions and Definitions

This section identifies the extensions and other documents that have been identified as likely to be needed to support the full set of MPLS-TP control plane requirements.

##### 4.4.1. MPLS-TE to MPLS-TP LSP Control Plane Interworking

While no interworking function is expected in the data-lane to support the interconnection of MPLS-TE and MPLS-TP networking, this is not the case for the control plane. MPLS-TE networks typically use LSP signaling based on [RFC3209] while MPLS-TP LSPs will be signaled using GMPLS RSVP-TE, i.e., [RFC3473]. The data plane of

[RFC5145] identifies a set of solutions that are aimed to aid in the interworking of MPLS-TE and GMPLS control planes. This work will serve as the foundation for a formal definition of MPLS to MPLS-TP control plane interworking.

##### 4.4.2. Associated Bidirectional LSPs

GMPLS signaling, [RFC3473], supports unidirectional, and co-routed bidirectional point-to-point LSPs. MPLS-TP also requires support for associated bidirectional point-to-point LSPs. Such support will require an extension or a formal definition of how the LSP endpoints supporting an associated bidirectional service will coordinate the two LSPs used to provide such a service. Per requirement 11, transit nodes that support an associated bidirectional service should be aware of the association of the LSPs used to support the service when both LSPs are supported on that transit node. There are several existing protocol mechanisms on which to base such support, including, but not limited to:

- o GMPLS calls, [RFC4974].
- o The ASSOCIATION object, [RFC4872].
- o The LSP\_TUNNEL\_INTERFACE\_ID object, [HIERARCHY-BIS].

##### 4.4.3. Asymmetric Bandwidth LSPs

[RFC5467] defines support for bidirectional LSPs which have different (asymmetric) bandwidth requirements for each direction. This RFC can be used to meet the related MPLS-TP technical requirement, but this RFC is currently an Experimental RFC. To fully satisfy the MPLS-TP requirement this document will need to become a Standards Track RFC.

#### 4.4.4. Recovery for P2MP LSPs

The definitions of P2MP, [RFC4875], and GMPLS recovery, [RFC4872] and [RFC4873], do not explicitly cover their interactions. MPLS-TP requires a formal definition of recovery techniques for P2MP LSPs. Such a formal definition will be based on existing RFCs and may not require any new protocol mechanisms, but nonetheless, must be documented.

#### 4.4.5. Test Traffic Control and other OAM functions

[CCAMP-OAM-FWK] and [CCAMP-OAM-EXT] are works in progress that extend the OAM related control capabilities of GMPLS. These extensions cover a portion, but not all OAM related control functions that have been identified in the context of MPLS-TP. As discussed above, the MPLS-TP control plane must support the selection of which (if any) OAM function(s) to use (including support to select experimental OAM functions) and what OAM functionality to run, including, continuity check (CC), connectivity verification (CV), packet loss and delay quantification, and diagnostic testing of a service. As OAM configuration is directly linked to data plane OAM, it is expected that [CCAMP-OAM-EXT] will evolve in parallel with the specification of data plane OAM functions. These documents do not yet cover the implications of SPMEs, including both dynamic creation and dynamic OAM function control.

#### 4.4.6. DiffServ Object usage in GMPLS

[RFC3270] and [RFC4124] define support for DiffServ enabled MPLS LSPs. While [RFC4124] references GMPLS signaling, there is no explicit discussion on the use of the DiffServ related objects in GMPLS signaling. A (possibly Informational) document on how GMPLS supports DiffServ LSPs is likely to prove useful in the context of MPLS-TP.

#### 4.4.7. Support for MPLS-TP LSP Identifiers

MPLS-TP uses two forms of LSP identifiers, see [TP-IDENTIFIERS]. One form is based on existing GMPLS fields. The other form is based on either the globally unique Attachment Interface Identifier (AII) defined in [RFC5003], or the M.1400 defined the ITU Carrier Code (ICC). Neither form is currently supported in GMPLS and such extensions will need to be documented.

#### 4.4.8. Support for MPLS-TP Maintenance Identifiers

MPLS-TP defines several forms of maintenance entity related identifiers. Both node unique and global forms are defined. Extensions will be required to GMPLS to support these identifiers. These extensions may be added to existing works in progress, such as [CCAMP-OAM-FWK] and [CCAMP-OAM-EXT], or may be defined in independent documents.

### 5. Pseudowires

#### 5.1. LDP Functions and Pseudowires

MPLS PWs are defined in [RFC3985] and [RFC5659], and provide for emulated services over an MPLS Packet Switched Network (PSN). Several types of PWs have been defined: (1) Ethernet PWs providing for Ethernet port or Ethernet VLAN transport over MPLS [RFC4448], (2) HDLC/PPP PW providing for HDLC/PPP leased line transport over MPLS [RFC4618], (3) ATM PWs [RFC4816], (4) Frame Relay PWs [RFC4619], and (5) circuit Emulation PWs [RFC4553].

Today's transport networks based on PDH, WDM, or SONET/SDH provide transport for PDH or SONET (e.g., ATM over SONET or Packet PPP over SONET) client signals with no payload awareness. Implementing PW capability allows for the use of an existing technology to substitute the TDM transport with packet based transport, using well-defined PW encapsulation methods for carrying various packet services over MPLS, and providing for potentially better bandwidth utilization.

There are two general classes of PWs: (1) Single-Segment Pseudowires (SS-PW) [RFC3985], and (2) Multi-segment Pseudowires (MS-PW) [RFC5659]. An MPLS-TP network domain may transparently transport a PW whose endpoints are within a client network. Alternatively, an MPLS-TP edge node may be the Terminating PE (T-PE) for a PW, performing adaptation from the native attachment circuit technology (e.g. Ethernet 802.1Q) to an MPLS PW which is then transported in an LSP over an MPLS-TP network. In this way, the PW is analogous to a transport channel in a TDM network and the LSP is equivalent to a container of multiple non-concatenated channels, albeit they are packet containers. An MPLS-TP network may also contain Switching PEs (S-PEs) for a multi-segment PW whereby the T-PEs may be at the edge of an MPLS-TP network or in a client network. In this latter case, a T-PE in a client network is a T-PE performing the adaptation of the native service to MPLS and an MPLS-TP network performs pseudowire switching.

The SS-PW signaling control plane is based on targeted LDP (T-LDP) with specific procedures defined in [RFC4447]. The MS-PW signaling control plane is also based on T-LDP as allowed for in [RFC5659], [SEGMENTED-PW] and [MS-PW-DYNAMIC]. An MPLS-TP network shall use the

same PW signaling protocols and procedures for placing SS-PWs and MS-PWs. This will leverage existing technology as well as facilitate interoperability with client networks with native attachment circuits or PW segments that are switched across an MPLS-TP network.

## 5.2. PW Control (LDP) and MPLS-TP Requirements Table

The following table shows how the MPLS-TP control plane requirements can be met using the existing LDP control plane for Pseudowires (targeted LDP). Areas where additional specifications are required are also identified. The table lists references based on the control plane requirements as identified and numbered above in section 2.

In the table below, several of the requirements shown are addressed - in part or in full - by the use of MPLS-TP LSPs to carry pseudowires. This is reflected by including "TP-LSPs" as a reference for those requirements. Section 5.3.2 provides additional context for the binding of PWs to TP-LSPs.

Req #	References
1	Generic requirement met by using Standards Track RFCs
2	[RFC3985], [RFC4447], Together with TP-LSPs (Sec. 4.3)
3	[RFC3985], [RFC4447]
4	Generic requirement met by using Standards Track RFCs
5	[RFC3985], [RFC4447], Together with TP-LSPs
6	[RFC3985], [RFC4447], [PW-P2MPR], [PW-P2MPE] + TP-LSPs
7	[RFC3985], [RFC4447], + TP-LSPs
8	[PW-P2MPR], [PW-P2MPE]
9	[RFC3985], end-node only involvement for PW
10	[RFC3985], proper vendor implementation
11	[RFC3985], end-node only involvement for PW
12-13	[RFC3985], [RFC4447], See Section 5.3.4
14	[RFC3985], [RFC4447]
15	[RFC4447], [RFC3478], proper vendor implementation
16	[RFC3985], [RFC4447]
17-18	[RFC3985], proper vendor implementation
19-26	[RFC3985], [RFC4447], [RFC5659], implementation
27	[RFC4448], [RFC4816], [RFC4618], [RFC4619], [RFC4553] [RFC4842], [RFC5287]
28	[RFC3985]
29-31	[RFC3985], [RFC4447]
32	[RFC3985], [RFC4447], [RFC5659], See Section 5.3.6.
33	[RFC4385], [RFC4447], [RFC5586]
34	[PW-P2MPR], [PW-P2MPE]
35	[RFC4863]
36-37	[RFC3985], [RFC4447], See Section 5.3.4
38	
39	Provided by TP-LSPs
40	[RFC3985], [RFC4447], + TP-LSPs
41	[RFC3478]
42-43	[RFC3985], [RFC4447]
44-45	[RFC3985], [RFC4447], + TP-LSPs - See Section 5.3.5
46	[RFC3985], [RFC4447], [RFC5659] + TP-LSPs
47	[RFC3985], [RFC4447], + TP-LSPs - See Section 5.3.3
48	[PW-RED], [PW-REDB]
49-50	[RFC3985], [RFC4447], + TP-LSPs, implementation
51-53	Provided by TP-LSPs, and Section 5.3.5
54-56	[RFC3985], [RFC4447], See Section 5.3.5
57	[PW-RED], [PW-REDB] revertive/non-revertive behavior is a local matter for PW
58-59	[PW-RED], [PW-REDB]
60-82	[RFC3985], [RFC4447], [PW-RED], [PW-REDB], Section 5.3.5
83-84	[RFC5085], [RFC5586], [RFC5885]
85-90	[RFC3985], [RFC4447], [PW-RED], [PW-REDB], Section 5.3.5
91-96	[RFC3985], [RFC4447], + TP-LSPs, implementation
97	[RFC4447], [MS-PW-DYNAMIC]
98	[RFC4447]
99 -	

100	Not Applicable to PW
101	[RFC4447]
102	[RFC3478]
103	[RFC3985], + TP-LSPs
104	Not Applicable to PW
105	[PW-OAM]
106	[PW-OAM]
107 -	
109	[RFC5085], [RFC5586], [RFC5885]
110	[RFC5085], [RFC5586], [RFC5885]
	fault reporting and protection triggering is a local matter for PW
111	[RFC5085], [RFC5586], [RFC5885]
	fault reporting and protection triggering is a local matter for PW
112	[RFC4447]
113	[RFC4447], [RFC5085], [RFC5586], [RFC5885]
114	[RFC5085], [RFC5586], [RFC5885]
115	[RFC5085], [RFC5586], [RFC5885]
116	path traversed by PW is determined by LSP path, see GMPLS and MPLS-TP Requirements Table, 4.3
117	[PW-RED], [PW-REDB], administrative control of redundant PW is a local matter at the PW head-end
118	[PW-RED], [PW-REDB], [RFC5085], [RFC5586], [RFC5885]
119	[RFC3985], [RFC4447], [PW-RED], [PW-REDB], Section 5.3.5
120	[RFC4447]
121 -	
126	[RFC5085], [RFC5586], [RFC5885]
127 -	
131	[PW-OAM]
132	Section 5.3.5
133	[PW-OAM]
134	[PW-OAM]
135	Section 5.3.5
136	[PW-OAM]
137	Not Applicable to PW
138	Not Applicable to PW
139	[RFC4447], [RFC5003], [MS-PW-DYNAMIC]
140 -	
144	[PW-OAM]

### 5.3. Anticipated MPLS-TP Related Extensions

The same control protocol and procedures will be reused as much as possible. However, when using PWs in MPLS-TP, a set of new requirements are defined which may require extensions of the existing control mechanisms. This section clarifies the areas where extensions are needed based on the PW Control Plane related requirements documented in [RFC5654].

See the table in the section above for a list of how requirements defined in [RFC5654] are expected to be addressed.

The baseline requirement for extensions to support transport applications is that any new mechanisms and capabilities must be able to interoperate with existing IETF MPLS [RFC3031] and IETF PWE3 [RFC3985] control and data planes where appropriate. Hence, extensions of the PW Control Plane must be in-line with the procedures defined in [RFC4447], [SEGMENTED-PW] and [MS-PW-DYNAMIC].

#### 5.3.1. Extensions to Support Out-of-Band PW Control

For MPLS-TP, it is required that the data and control planes can be both logically and physically separated. That is, the PW Control Plane must be able to operate out-of-band (OOB). This separation ensures, among other things, that in the case of control plane failures the data plane is not affected and can continue to operate normally. This was not a design requirement for the current PW Control Plane. However, due to the PW concept, i.e., PWs are connecting logical entities ('forwarders'), and the operation of the PW control protocol, i.e., only edge PE nodes (T-PE, S-PE) take part in the signaling exchanges: moving T-LDP out-of-band seems to be, theoretically, a straightforward exercise.

In fact, as a strictly local matter, ensuring that targeted LDP (T-LDP) uses out-of-band signaling requires only that the local implementation is configured in such a way that reachability for a target LSR address is via the out-of-band channel.

More precisely, if IP addressing is used in the MPLS-TP control plane then T-LDP addressing can be maintained, although all addresses will refer to control plane entities. Both, the PWid FEC and Generalized PWid FEC Elements can possibly be used in an OOB case as well. (Detailed evaluation is outside the scope of this document). The PW Label allocation and exchange mechanisms should be reused without change.

#### 5.3.2. Support for Explicit Control of PW-to-LSP Binding

Binding a PW to an LSP, or PW segments to LSPs is left to nodes acting as T-PEs and S-PEs or a control plane entity that may be the same one signaling the PW. However, an extension of the PW signaling protocol is required to allow the LSR at signal initiation end to inform the targeted LSR (at the signal termination end) which LSP the resulting PW is to be bound to, in the event that more than one such LSP exists and the choice of LSPs is important to the service being setup (for example, if the service requires co-routed bidirectional paths). This is also particularly important to support transport path (symmetric and asymmetric) bandwidth requirements.

If the control plane is physically separated from the forwarder, the control plane must be able to program the forwarders with necessary information.

For transport services, it may be required that bidirectional traffic follows congruent paths. Currently, each direction of a PW or a PW segment is bound to a unidirectional LSP that extends between two T-PEs, S-PEs, or a T-PE and an S-PE. The unidirectional LSPs in both directions are not required to follow congruent paths, and therefore both directions of a PW may not follow congruent paths, i.e., they are associated bidirectional paths. The only requirement in [RFC5659] is that a PW or a PW segment shares the same T-PEs in both directions, and same S-PEs in both directions.

MPLS-TP imposes new requirements on the PW Control Plane, in requiring that both end points map the PW or PW segment to the same transport path for the case where this is an objective of the service. When a bidirectional LSP is selected on one end to transport the PW, a mechanism is needed that signals to the remote end which LSP has been selected locally to transport the PW. This would be accomplished by adding a new TLV to PW signaling.

Note that this coincides with the gap identified for OOB support: a new mechanism is needed to allow explicit binding of a PW to the supporting transport LSP.

The case of unidirectional transport paths may also require additional protocol mechanisms as today's PWs are always bidirectional. One potential approach for providing a unidirectional PW based transport path is for the PW to associate different (asymmetric) bandwidths in each direction, with a zero or minimal bandwidth for the return path. This approach is consistent with Section 3.8.2 of [RFC5921] but does not address P2MP paths.

#### 5.3.3. Support for Dynamic Transfer of PW Control/Ownership

In order to satisfy requirement 47 (as defined in section 2) it will be necessary to specify methods for transfer of PW ownership from the management to the control plane (and vice versa).

#### 5.3.4. Interoperable Support for PW/LSP Resource Allocation

Transport applications may require resource guarantees. For such transport LSPs, resource reservation mechanisms are provided via RSVP-TE and the use of DiffServ. If multiple PWs are multiplexed into the same transport LSP resources, contention may occur. However, local policy at PEs should ensure proper resource sharing among PWs mapped into a resource guaranteed LSP. In the case of MS-PWs, signaling carries the PW traffic parameters [MS-PW-DYNAMIC] to enable

admission control of a PW segment over a resource-guaranteed LSP.

In conjunction with explicit PW-to-LSP binding, existing mechanisms may be sufficient, however this needs to be verified in detailed evaluation.

#### 5.3.5. Support for PW Protection and PW OAM Configuration

Many of the requirements listed in section 2 are intended to support connectivity and performance monitoring (grouped together as OAM) and protection conformant with the transport services model.

In general, protection of MPLS-TP transported services is provided by way of protection of transport LSPs. PW protection requires that mechanisms be defined to support redundant Pseudowires, including a mechanism already described above for associating such Pseudowires with specific protected ("working" and "protection") LSPs. Also required are definitions of local protection control functions, to include test/verification operations, and protection status signals needed to ensure that PW termination points are in agreement as to which of a set of redundant Pseudowires are in use for which transport services at any given point in time.

Much of this work is currently being done in drafts [PW-RED] and [PW-REDB] that define - respectively - how to establish redundant Pseudowires and how to indicate which is in use. Additional work may be required.

Protection switching may be triggered manually by the operator, or as a result of loss of connectivity (detected using the mechanisms of [RFC5085] and [RFC5586]), or service degradation (detected using mechanisms yet to be defined).

Automated protection switching is just one of the functions for which a transport service require OAM. OAM is generally referred to as either "proactive" or "on-demand", where the distinction is whether a specific OAM tool is being used continuously over time (for the purpose of detecting a need for protection switching, for example) or is only used - either a limited number of times, or over a short period of time - when explicitly enabled (for diagnostics, for example).

PW OAM currently consists of connectivity verification defined by [RFC5085]. Work is currently in progress to extend PW OAM to include bidirectional forwarding detection (BFD) in [RFC5885], and work has begun on extending BFD to include performance related monitor functions.

#### 5.3.6. Client Layer and Cross-Provider Interfaces to PW Control

Additional work is likely to be required to define consistent access by a client layer network, as well as between provider networks, to control information available to each type of network, for example, about the topology of an MS-PW. This information may be required by the client layer network in order to provide hints that may help to avoid establishment of fate-sharing alternate paths. Such work will need to fit within the ASON architecture, see requirement 39 above.

#### 5.4. ASON Architecture Considerations

MPLS-TP PWs are always transported using LSPs, and these LSP will either have been statically provisioned or signaled using GMPLS.

For LSPs signaled using the MPLS-TP LSP control plane (GMPLS), conformance with the ASON architecture is as described in Section 1.2 ("Basic Approach"), bullet 4, of this framework document.

As discussed above in Section 5.3, there are anticipated extensions in the following areas that may be related to ASON architecture:

- PW-to-LSP binding (Section 5.3.2)
- PW/LSP resource allocation (Section 5.3.4)
- PW protection and OAM configuration (Section 5.3.5)
- Client layer Interfaces for PW control (Section 5.3.6)

This work is expected to be consistent with ASON architecture and may require additional specification in order to achieve this goal.

#### 6. Security Considerations

This document primarily describes how existing mechanisms can be used to meet the MPLS-TP control plane requirements. The documents that describe each mechanism contain their own security considerations sections. For a general discussion on MPLS and GMPLS related security issues, see the MPLS/GMPLS security framework [RFC5920].

This document also identifies a number of needed control plane extensions. It is expected that the documents that define such extensions will also include any appropriate security considerations.

## 7. IANA Considerations

There are no new IANA considerations introduced by this document.

## 8. Acknowledgments

The authors would like to acknowledge the contributions of Yannick Brehon, Diego Caviglia, Nic Neate, Dave Mcdysan, Dan Frost, and Eric Osborne to this work. We also thank Dan Frost in his help responding to last call comments.

## 9. References

### 9.1. Normative References

- [RFC2210] Wroclawski, J., "The Use of RSVP with Integrated Services", RFC 2210, September 1997.
- [RFC2211] Wroclawski, J., "Specification of the Controlled Load Quality of Service", RFC 2211, September 1997.
- [RFC2212] Shenker, S., Partridge, C., and R Guerin, "Specification of Guaranteed Quality of Service", RFC 2212, September 1997.
- [RFC3031] Rosen, E., Viswanathan, A., Callon, R., "Multiprotocol Label Switching Architecture", RFC 3031, January 2001.
- [RFC3209] Awduche, D., Berger, L., Gan, D., Li, T., Srinivasan, V., and G. Swallow, "RSVP-TE: Extensions to RSVP for LSP Tunnels", RFC 3209, December 2001.
- [RFC3471] Berger, L., "Generalized Multi-Protocol Label Switching (GMPLS) Signaling Functional Description", RFC 3471, January 2003.
- [RFC3473] Berger, L. Ed., "Generalized Multi-Protocol Label Switching (GMPLS) Signaling Resource ReserVation Protocol-Traffic Engineering (RSVP-TE) Extensions", RFC 3473, January 2003.
- [RFC3478] Leelanivas, M, et al, "Graceful Restart Mechanism for Label Distribution Protocol", RFC 3478, February 2003.
- [RFC3630] Katz, D., Kompella, K., and D. Yeung, "Traffic Engineering (TE) Extensions to OSPF Version 2", RFC 3630, September 2003.

- [RFC4124] Le Faucheur, F., Ed. "Protocol Extensions for Support of Diffserv-aware MPLS Traffic Engineering", RFC 4124, June 2005.
- [RFC4202] Kompella, K. and Y. Rekhter, "Routing Extensions in Support of Generalized Multi-Protocol Label Switching (GMPLS)", RFC 4202, October 2005.
- [RFC4203] Kompella, K. and Y. Rekhter, "OSPF Extensions in Support of Generalized Multi-Protocol Label Switching (GMPLS)", RFC 4203, October 2005.
- [RFC4206] Kompella, K. and Y. Rekhter, "Label Switched Paths (LSP) Hierarchy with Generalized Multi-Protocol Label Switching (GMPLS) Traffic Engineering (TE)", RFC 4206, October 2005.
- [RFC4385] Bryant, S., et al, "Pseudowire Emulation Edge-to-Edge (PWE3) Control Word for Use over an MPLS PSN", RFC 4385, February 2006.
- [RFC4447] Martini, L., Ed., "Pseudowire Setup and Maintenance Using the Label Distribution Protocol (LDP)", RFC 4447, April 2006.
- [RFC4448] Martini, L., Ed., "Encapsulation Methods for Transport Ethernet over MPLS Network", RFC 4448, April 2006.
- [RFC4842] Malis, A., et al, "Synchronous Optical Network/ Synchronous Digital Hierarchy (SONET/SDH) Circuit Emulation over Packet (CEP)", RFC 4842, April 2007.
- [RFC4863] Martini, L. and G. Swallow, "Wildcard Pseudowire Type", RFC 4863, May 2007.
- [RFC4872] Lang, J., Rekhter, Y., and Papadimitriou, D., "RSVP-TE Extensions in Support of End-to-End Generalized Multi-Protocol Label Switching (GMPLS) Recovery", RFC 4872, May 2007.
- [RFC4873] Berger, L., Bryskin, I., Papadimitriou, D., Farrel, A., "GMPLS Segment Recovery", RFC 4873, May 2007.
- [RFC4929] Andersson, L. and A. Farrel, "Change Process for Multiprotocol Label Switching (MPLS) and Generalized MPLS (GMPLS) Protocols and Procedures", BCP 129, RFC 4929, June 2007.

- [RFC4974] Papadimitriou, D., Farrel, A., "Generalized MPLS (GMPLS) RSVP-TE Signaling Extensions in Support of Calls", RFC 4974, August 2007.
- [RFC5063] Satyanarayana, A., Ed., "Extensions to GMPLS Resource Reservation Protocol (RSVP) Graceful Restart", RFC 5063, September 2007.
- [RFC5287] Vainshtein, A. and Y. Stein, "Control Protocol Extensions for the Setup of Time-Division Multiplexing (TDM) Pseudowires in MPLS Networks", RFC 5287, August 2008.
- [RFC5305] Smit, H. and T. Li, "IS-IS Extensions for Traffic Engineering", RFC 5305, October 2008.
- [RFC5307] Kompella, K. and Rekhter, Y., "IS-IS Extensions in Support of Generalized Multi-Protocol Label Switching (GMPLS)", RFC 5307, October 2008.
- [RFC5316] Chen, M., Zhang, R., and Duan, X., "ISIS Extensions in Support of Inter-Autonomous System (AS) MPLS and GMPLS Traffic Engineering", RFC 5316, December 2008.
- [RFC5392] Chen, M., Zhang, R., and Duan, X., "OSPF Extensions in Support of Inter-Autonomous System (AS) MPLS and GMPLS Traffic Engineering", RFC 5392, January 2009.
- [RFC5151] Farrel, A., Ed., "Inter-Domain MPLS and GMPLS Traffic Engineering -- Resource Reservation Protocol-Traffic Engineering (RSVP-TE) Extensions", RFC 5151, February 2008.
- [RFC5654] Niven-Jenkins, B., et al, "Requirements of an MPLS Transport Profile", RFC 5654, September 2009.
- [RFC5467] Berger, L., et al, "GMPLS Asymmetric Bandwidth Bidirectional Label Switched Paths (LSPs)", RFC 5467, March 2009.
- [RFC5586] Bocci, M., et al, "MPLS Generic Associated Channel", RFC 5586, June 2009.
- [RFC5860] Vigoureux, M., Ward, D., Betts, M., "Requirements for Operations, Administration, and Maintenance (OAM) in MPLS Transport Networks", RFC 5860, May 2010.
- [RFC5921] Bocci, M., Bryant, S., Frost, D., Levrau, L., Berger, L., "A Framework for MPLS in Transport Networks", RFC 5921, July 2010.

- [RFC5960] Frost, D., Bryant, S., Bocci, M., "MPLS Transport Profile Data Plane Architecture", RFC 5960, August 2010.
- [TP-IDENTIFIERS] Bocci, M., Swallow, G., "MPLS-TP Identifiers", work in progress, draft-ietf-mpls-tp-identifiers.
- [TP-OAM] Busi, I., Ed., Niven-Jenkins, B., Ed., "MPLS-TP OAM Framework and Overview", work in progress, draft-ietf-mpls-tp-oam-framework.
- [TP-SURVIVE] Sprecher, N., et al., "Multiprotocol Label Switching Transport Profile Survivability Framework", work in progress, draft-ietf-mpls-tp-survive-fwk.

## 9.2. Informative References

- [CCAMP-OAM-FWK] A. Takacs, D. Fedyk, and J. He, "OAM Configuration Framework and Requirements for GMPLS RSVP-TE", work in progress, draft-ietf-ccamp-oam-configuration-fwk.
- [CCAMP-OAM-EXT] Bellagamba, E., et.al., "RSVP-TE Extensions for MPLS-TP OAM Configuration", work in progress, draft-bellagamba-ccamp-rsvp-te-mpls-tp-oam-ext.
- [GMPLS-PS] Takacs, A., et al, "GMPLS RSVP-TE Recovery Extension for data plane initiated reversion and protection timer signalling", work in progress, draft-takacs-ccamp-revertive-ps.
- [HIERARCHY-BIS] Shiimoto, K, Ed., Farrel, A, Ed., "Procedures for Dynamically Signaled Hierarchical Label Switched Paths", work in progress, draft-ietf-ccamp-lsp-hierarchy-bis.
- [TE-MIB] T Otani, et.al., "Traffic Engineering Database Management Information Base in support of MPLS-TE/GMPLS", work in progress, draft-ietf-ccamp-gmpls-td-mib.
- [MS-PW-DYNAMIC] L. Martini, M Bocci, and F Balus "Dynamic Placement of Multi Segment Pseudo Wires", work in progress, draft-ietf-pwe3-dynamic-ms-pw.
- [ITU.G8080.2006] International Telecommunications Union, "Architecture for the automatically switched optical network (ASON)", ITU- T Recommendation G.8080, June 2006.

- [ITU.G8080.2008] International Telecommunications Union,  
"Architecture for the automatically switched  
optical network (ASON) Amendment 1", ITU-T  
Recommendation G.8080 Amendment 1, March 2008.
- [NO-PHP] Ali, z., et al, "Non PHP Behavior and out-of-band mapping  
for RSVP-TE LSPs", work in progress,  
draft-ietf-mpls-rsvp-te-no-php-oob-mapping
- [PW-RED] Muley, P., et al, "Pseudowire (PW) Redundancy", work in  
progress, draft-ietf-pwe3-redundancy.
- [PW-REDB] Muley, P., et al, "Preferential Forwarding Status bit  
definition", work in progress,  
draft-ietf-pwe3-redundancy-bit.
- [PW-OAM] Zhang, F., et al, "LDP Extensions for MPLS-TP PW OAM  
configuration", work in progress,  
draft-zhang-mpls-tp-pw-oam-config.
- [PW-P2MPE] Aggarwal, R. and F. Jounay, "Point-to-Multipoint  
Pseudo-Wire Encapsulation", work in progress,  
draft-raggarwa-pwe3-p2mp-pw-encaps.
- [PW-P2MPR] Jounay, F., et al, "Requirements for  
Point-to-Multipoint Pseudowire", work in progress,  
draft-ietf-pwe3-p2mp-pw-requirements.
- [RFC3270] Le Faucheur, F., et al, "Multi-Protocol Label Switching  
(MPLS) Support of Differentiated Services", RFC 3270,  
May 2002.
- [RFC3472] Ashwood-Smith, P., Ed, Berger, L. Ed., "Generalized  
Multi-Protocol Label Switching (GMPLS) Signaling  
Constraint-based Routed Label Distribution Protocol  
(CR-LDP) Extensions", RFC 3472, January 2003.
- [RFC3477] Kompella, K., Rekhter, Y., "Signalling Unnumbered Links  
in Resource ReSerVation Protocol - Traffic Engineering  
(RSVP-TE)", RFC 3477, January 2003.
- [RFC3478] Leelanivas, M., Rekhter, Y., Aggarwal, R., "Graceful  
Restart Mechanism for Label Distribution Protocol", RFC  
3478, February 2003.
- [RFC3812] Srinivasan, C., Viswanathan, A., and T. Nadeau,  
"Multiprotocol Label Switching (MPLS) Traffic  
Engineering (TE) Management Information Base (MIB)", RFC  
3812, June 2004.

- [RFC3813] Srinivasan, C., Viswanathan, A., and T. Nadeau, "Multiprotocol Label Switching (MPLS) Label Switching (LSR) Router Management Information Base (MIB)", RFC 3813, June 2004.
- [RFC3945] Mannie, E., "Generalized Multi-Protocol Label Switching (GMPLS) Architecture", RFC 3945, October 2004.
- [RFC3985] Bryant, S. and P. Pate, "Pseudowire Emulation Edge-to-Edge (PWE3) Architecture", RFC 3985, March 2005.
- [RFC4139] Papadimitriou, D., et al, "Requirements for Generalized MPLS (GMPLS) Signaling Usage and Extensions for Automatically Switched Optical Network (ASON)", RFC4139, July 2005.
- [RFC4201] Kompella, K., Rekhter, Y., Berger, L., "Link Bundling in MPLS Traffic Engineering (TE)", RFC 4201, October 2005.
- [RFC4208] Swallow, G., Drake, J., Ishimatsu, H., and Rekhter, Y., "Generalized Multi-Protocol Label Switching (GMPLS) User-Network Interface (UNI) : Resource Reservation Protocol-Traffic Engineering (RSVP-TE) Support for the Overlay Model", RFC 4208, October 2005.
- [RFC4258] Brungard, D., et al, "Requirements for Generalized Multi-Protocol Label Switching (GMPLS) Routing for the Automatically Switched Optical Network (ASON)", RFC4258, November 2005.
- [RFC4379] Kompella, K. and G. Swallow, "Detecting Multi-Protocol Label Switched (MPLS) Data Plane Failures", RFC 4379, February 2006.
- [RFC4426] Lang, J., Rajagopalan B., and D.Papadimitriou, Editors, "Generalized Multiprotocol Label Switching (GMPLS) Recovery Functional Specification", RFC 4426, March 2006.
- [RFC4427] Mannie, E., Papadimitriou, D., "Recovery (Protection and Restoration) Terminology for Generalized Multi-Protocol Label Switching (GMPLS)", RFC4427, March 2006.
- [RFC4553] Vainshtein, A., Ed., and Stein, YJ., Ed., "Structure-Agnostic Time Division Multiplexing (TDM) over Packet (SAToP)", RFC 4553, June 2006.

- [RFC4618] Martini, L., Rosen, E., Heron, G., and Malis, A.,  
"Encapsulation Methods for Transport of PPP/High- Level  
Data Link Control (HDLC) over MPLS Networks", RFC 4618,  
September 2006.
- [RFC4619] Martini, L., Ed., Kawa, C., Ed., and Malis, A., Ed.,  
"Encapsulation Methods for Transport of Frame Relay over  
Multiprotocol Label Switching (MPLS) Networks",  
September 2006.
- [RFC4655] Farrel, A., Vasseur, J.-P., Ash, J., "A Path Computation  
Element (PCE)-Based Architecture", RFC 4655, August  
2006.
- [RFC4783] Berger, L., Ed., "GMPLS - Communication of Alarm  
Information", RFC 4763, December 2006.
- [RFC4802] T. D. Nadeu and A. Farrel, "Generalized Multiprotocol  
Label Switching (GMPLS) Traffic Engineering Management  
Information Base", RFC 4802, February 2007.
- [RFC4803] T. D. Nadeu and A. Farrel, "Generalized Multiprotocol  
Label Switching (GMPLS) Label Switching Router (LSR)  
Management Information Base", RFC 4803, February 2007.
- [RFC4816] Malis, A., Martini, L., Brayley, J., and Walsh, T.,  
"Pseudowire Emulation Edge-to-Edge (PWE3) Asynchronous  
Transfer Mode (ATM) Transparent Cell Transport Service",  
RFC 4816, February 2007.
- [RFC4875] Aggarwal, R., Papadimitriou, D., Yasukawa, S.,  
"Extensions to Resource Reservation Protocol - Traffic  
Engineering (RSVP-TE) for Point-to-Multipoint TE Label  
Switched Paths (LSPs)", RFC 4875, May 2007.
- [RFC5003] Metz, C., Martini, L., Balus, F., Sugimoto, J.,  
"Attachment Individual Identifier (AII) Types for  
Aggregation", RFC 5003, September 2007.
- [RFC5036] Andersson, L., I. Minei and B. Thomas, Editors, "LDP  
Specification", RFC 5036, October 2007.
- [RFC5085] Nadeau, T. and C. Pignataro, "Pseudowire Virtual Circuit  
Connectivity Verification (VCCV) : A Control Channel for  
Pseudowires", RFC 5085, December 2007.
- [RFC5145] Shiimoto, K., "Framework for MPLS-TE to GMPLS  
Migration", RFC 5145, March 2008.

- [RFC5440] Vasseur, JP., Le, JL., "Path Computation Element (PCE) Communication Protocol (PCEP)", RFC 5440, March 2009.
- [RFC5493] Caviglia, D., et al, "Requirements for the Conversion between Permanent Connections and Switched Connections in a Generalized Multiprotocol Label Switching (GMPLS) Network", RFC 5493, April 2009.
- [RFC5659] Bocci, M., and Bryant, B., "An Architecture for Multi-Segment Pseudowire Emulation Edge-to-Edge", RFC 5659, October 2009.
- [RFC5787] Papadimitriou, D., "OSPFv2 Routing Protocols Extensions for ASON Routing", RFC 5787, March 2010.
- [RFC5852] Caviglia, D., Ceccarelli, D., Bramanti, D., Li, D., Bardalai, S., "RSVP-TE Signaling Extension for LSP Handover from the Management Plane to the Control Plane in a GMPLS-Enabled Transport Network", RFC 5852, April 2010.
- [RFC5884] Aggarwal, R., Kompella, K., Nadeau, T., and G. Swallow, "Bidirectional Forwarding Detection (BFD) For MPLS Label Switched Paths (LSPs)", RFC 5884, June 2010.
- [RFC5885] Nadeau, T. and C. Pignataro, "Bidirectional Forwarding Detection (BFD) for the Pseudowire Virtual Circuit Connectivity Verification (VCCV)", RFC 5885, June 2010.
- [RFC5920] Fang, L., "Security Framework for MPLS and GMPLS Networks", RFC 5920, July 2010.
- [RFC6001] Papadimitriou, D., et al, "Generalized Multi-Protocol Label Switching (GMPLS) Protocol Extensions for Multi-Layer and Multi-Region Networks (MLN/MRN)", RFC 6001, October 2010.
- [SEGMENTED-PW] Martini, L., Nadeau, T., and Duckett M., "Segmented Pseudowire", work in progress, draft-ietf-pwe3-segmented-pw.
- [TP-P2MP-FWK] D. Frost, M. Bocci, and L. Berger, "A Framework for Point-to-Multipoint MPLS in Transport Networks", draft-fbb-mpls-tp-p2mp-framework.
- [TP-RING] Weingarten, Y., Ed., "MPLS-TP Ring Protection", work in progress, draft-weingarten-mpls-tp-ring-protection.

## 10. Authors' Addresses

Loa Andersson (editor)  
Ericsson  
Phone: +46 10 717 52 13  
Email: loa.andersson@ericsson.com

Lou Berger (editor)  
LabN Consulting, L.L.C.  
Phone: +1-301-468-9228  
Email: lberger@labn.net

Luyuan Fang (editor)  
Cisco Systems, Inc.  
300 Beaver Brook Road  
Boxborough, MA 01719  
USA  
Email: lufang@cisco.com

Nabil Bitar (editor)  
Verizon  
117 West Street  
Waltham, MA 02451  
Email: nabil.n.bitar@verizon.com

Eric Gray (editor)  
Ericsson  
900 Chelmsford Street  
Lowell, MA, 01851  
Phone: +1 978 275 7470  
Email: Eric.Gray@Ericsson.com

Attila Takacs  
Ericsson  
1. Laborc u.  
Budapest, HUNGARY 1037  
Email: attila.takacs@ericsson.com

Martin Vigoureux  
Alcatel-Lucent  
Email: martin.vigoureux@alcatel-lucent.fr

Elisa Bellagamba  
Ericsson  
Farogatan, 6  
164 40, Kista, Stockholm, SWEDEN  
Email: elisa.bellagamba@ericsson.com

Generated on: Fri, Oct 15, 2010 2:54:52 PM

Network working group  
Internet Draft  
Category: Standards Track  
Created: October 25, 2010  
Expires: April 2011

M. Chen (Ed.)  
Huawei Technologies Co., Ltd  
N. So (Ed.)  
Verizon

## Return Path Specified LSP Ping

draft-ietf-mpls-return-path-specified-lsp-ping-01.txt

### Abstract

This document defines extensions to the failure-detection protocol for Multiprotocol Label Switching (MPLS) Label Switched Paths (LSPs) known as "LSP Ping" that allow selection of the LSP to use for the echo reply return path. Enforcing a specific return path can be used to verify bidirectional connectivity and also increase LSP ping robustness. It may also be used by Bidirectional Forwarding Detection (BFD) for MPLS bootstrap signaling thereby making BFD for MPLS more robust.

### 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 March 18, 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 BSD License.

## Conventions used in 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 RFC-2119 [RFC2119].

## Table of Contents

1. Introduction .....	3
2. Problem Statements and Solution Overview .....	3
2.1. Limitations of Existing Mechanisms for Bidirectional LSPs .....	4
2.2. Limitations of Existing Mechanisms for Handling Unreliable Return Paths .....	4
3. Extensions .....	5
3.1. Reply Via Specified Path mode .....	6
3.2. Reply Path (RP) TLV .....	6
3.3. RP TLV sub-TLVs.....	8
3.3.1. IPv4 RSVP Tunnel sub-TLV .....	9
3.3.2. IPv6 RSVP Tunnel sub-TLV .....	11
3.3.3. RP TC sub-TLV .....	12
4. Theory of Operation .....	12
4.1. Sending an Echo Request .....	13
4.2. Receiving an Echo Request .....	13
4.3. Sending an Echo Reply .....	14
4.4. Receiving an Echo Reply .....	15
5. Security Considerations .....	16
6. IANA Considerations .....	16
6.1. Reply mode .....	16
6.2. RP TLV .....	16
6.3. Sub-TLVs for RP TLV .....	17

7. Contributors .....	17
8. Acknowledgments .....	18
9. References .....	18
9.1. Normative References .....	18
9.2. Informative References .....	19
Authors' Addresses .....	20

## 1. Introduction

This document defines extensions to the failure-detection protocol for Multiprotocol Label Switching (MPLS) Label Switched Paths (LSPs) known as "LSP Ping" [RFC4379] that can be used to specify the return paths for the echo reply message, increasing the robustness of LSP Ping, reducing the opportunity for error, and improving the reliability of the echo reply message. A new reply mode, which is referred to as "Reply via specified path", is added and a new Type-Length-Value (TLV), which is referred to as Reply Path (RP) TLV, is defined in this memo.

With the extensions described in this document, a bidirectional LSP and a pair of unidirectional LSPs (one for each direction) could both be tested with a single operational action, hence providing better control plane scalability. The defined extensions can also be utilized for creating a single Bidirectional Forwarding Detection (BFD) [BFD], [BFD-MPLS] session for a bidirectional LSP or for a pair of unidirectional LSPs (one for each direction).

In this document, term bidirectional LSP includes the co-routed bidirectional LSP defined in [RFC3945] and the associated bidirectional LSP that is constructed from a pair of unidirectional LSPs (one for each direction), and which are associated with one another at the LSP's ingress/egress points [RFC5654].

## 2. Problem Statements and Solution Overview

MPLS LSP Ping is defined in [RFC4379]. It can be used to detect data path failures in all MPLS LSPs, and was originally designed for unidirectional LSPs.

LSP are increasingly being deployed to provide bidirectional services. The co-routed bidirectional LSP is defined in [RFC3471] and [RFC3473], and the associated bidirectional LSP is defined in [RFC5654]. With the deployment of such services, operators have a desire to test both directions of a bidirectional LSP in a single operation.

Additionally, when testing a single direction of an LSP (either a unidirectional LSP, or a single direction of a bidirectional LSP) using LSP Ping, the validity of the result may be affected by the success of delivering the echo reply message. Failure to exchange these messages between the egress Label Switching Router (LSR) and the ingress LSR can lead to false negatives where the LSP under test is reported as "down" even though it is functioning correctly.

## 2.1. Limitations of Existing Mechanisms for Bidirectional LSPs

With the existing LSP Ping mechanisms as defined in [RFC4379], operators have to enable LSP detection on each of the two ends of a bidirectional LSP independently. This not only doubles the workload for the operators, but may also bring additional difficulties when checking the backward direction of the LSP under the following conditions:

1. The LSR that the operator logged on to perform the checking operations might not have out-of-band connectivity to the LSR at the far end of the LSP. That can mean it is not possible to check the return direction of a bidirectional LSP in a single operation - the operator must log on to the LSR at the other end of the LSP to test the return direction.
2. The LSP being tested might be an inter-domain/inter-AS LSP where the operator of one domain/AS may have no right to log on to the LSR at the other end of the LSP since this LSR resides in another domain/AS. That can make it completely impossible for the operator to check the return direction of a bidirectional LSP.

Associated bidirectional LSPs have the same issues as those listed for co-routed bidirectional LSPs.

This document defines a mechanism to allow the operator to request that both directions of a bidirectional LSP be tested by a single LSP Ping message exchange.

## 2.2. Limitations of Existing Mechanisms for Handling Unreliable Return Paths

[RFC4379] defines 4 reply modes:

1. Do not reply
2. Reply via an IPv4/IPv6 UDP packet
3. Reply via an IPv4/IPv6 UDP packet with Router Alert
4. Reply via application level control channel.

Obviously, the issue of the reliability of the return path for an echo reply message does not apply in the first of these cases.

[RFC4379] states that the third mode may be used when the IP return path is deemed unreliable. This mode of operation requires that all intermediate nodes must support the Router Alert option and must understand and know how to forward MPLS echo replies.

This is a rigorous requirement in deployed IP/MPLS networks especially since the return path may be through legacy IP-only routers. Furthermore, for inter-domain LSPs, the use of the Router Alert option may encounter significant issues at domain boundaries where the option is usually stripped from all packets. Thus, the use of this mode may itself introduce issues that lead to the echo reply messages not being delivered.

And in any case, the use modes 2 or 3 cannot guarantee the delivery of echo responses through an IP network that is fundamentally unreliable. The failure to deliver echo response messages can lead to false negatives making it appear that the LSP has failed.

Allowing the ingress LSR to control the path used for echo reply messages, and in particular forcing those messages to use an LSP rather than being sent through the IP network, enables an operator to apply an extra level of deterministic process to the LSP Ping test.

This document defines extensions to LSP Ping that can be used to specify the return paths of the echo reply message in an LSP echo request message.

### 3. Extensions

LSP Ping defined in [RFC4379] is carried out by sending an echo request message. It carries the Forwarding Equivalence Class (FEC) information of the tested LSP which indicates which MPLS path is being verified, along the same data path as other normal data packets belonging to the FEC.

LSP Ping [RFC4379] defines four reply modes that are used to direct the egress LSR in how to send back an echo reply. This document

defines a new reply mode, the Reply Via Specified Path mode. This new mode is used to direct the egress LSR of the tested LSP to send the echo reply message back along the path specified in the echo request message.

In addition, a new TLV, the Reply Path (RP) TLV, is defined in this document. The RP TLV consists of one or more sub-TLVs that can be used to carry the specified return path information to be used by the echo reply message.

### 3.1. Reply Via Specified Path mode

A new reply mode is defined to be carried in the Reply Mode field of the LSP Ping echo request message.

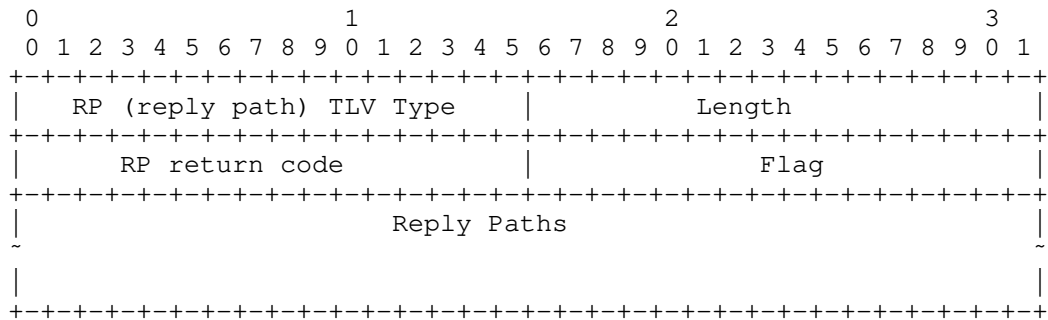
The recommended value of the Reply Via Specified Path mode is 5 (This is to be confirmed by the IANA).

Value	Meaning
-----	-----
5	Reply via specified path

The Reply Via Specified Path mode is used to notify the remote LSR receiving the LSP Ping echo request message to send back the echo reply message along the specified paths carried in the Reply Path TLV.

### 3.2. Reply Path (RP) TLV

The Reply Path (RP) TLV is optionally included in an echo request message. It carries the specified return paths that the echo reply message is required to follow. The format of RP TLV is as follows:



RP TLV Type field is 2 octets in length, and the type value is TBD by IANA.

The Length field is 2 octets in length. It defines the length in octets of the RP return code, Flag and Reply Paths fields.

RP return code is 2 octets in length. It is defined for the egress LSR of the forward LSP to report the results of RP TLV processing and return path selection. When sending echo request, these codes MUST set to zero. RP return code only used when sending echo reply, and it will be ignored when processing echo request message. There are 8 RP return codes defined in this document:

Value	meaning
0	No return code
1	Malformed RP TLV was received
2	One or more of the sub-TLVs in RP TLV was not understood
3	The echo reply was sent successfully using the specified RP
4	The specified RP was not found, the echo reply was sent via other LSP
5	The specified RP was not found, the echo reply was sent via IP path
6	The Reply mode in echo request was not set to 5 (replay via specified path) although RP TLV exists
7	RP TLV was missing in echo request

Flag field is also 2 octets in length, it is used to notify the egress how to process the Reply Paths field when performing return path selection. The Flag field is a bit vector and has following format:

```

+---+---+---+---+---+---+---+---+---+---+
|           MUST be zero           |A|B|E|
+---+---+---+---+---+---+---+---+---+---+

```

A (Alternative path): the egress LSR MUST select a non-default path as the return path. This is very useful when reverse default path problems are suspected which can be confirmed when the echo reply is forced to follow a non-default return path. If A bit is set, there is no need to carry any specific reply path sub-TLVs.

B (Bidirectional): the return path is required to follow the reverse direction of the tested bidirectional LSP.

E (Exclude): the return path is required to exclude the paths that are identified by the reply path sub-TLVs carried in the Reply Paths field. This is very useful when one or more previous LSP Ping attempts failed. By setting this E bit and carrying the previous failed reply path sub-TLVs, a new LSP Ping echo request could be used to help the egress LSR to select another candidate path when sending echo reply message.

A bit MUST not be set when any one of other two bits (B bit and E bit) set.

The Reply Paths field is variable in length. It has several nested sub-TLVs that describe the specified paths the echo reply message is required to follow. When the Reply Mode field is set to "Reply via specified path" in an LSP echo request message, the RP TLV MUST be present.

### 3.3. RP TLV sub-TLVs

Each of the FEC sub-TLVs defined in [RFC4379] is applicable to be a sub-TLV for inclusion in the RP TLV for expressing a specific return path.

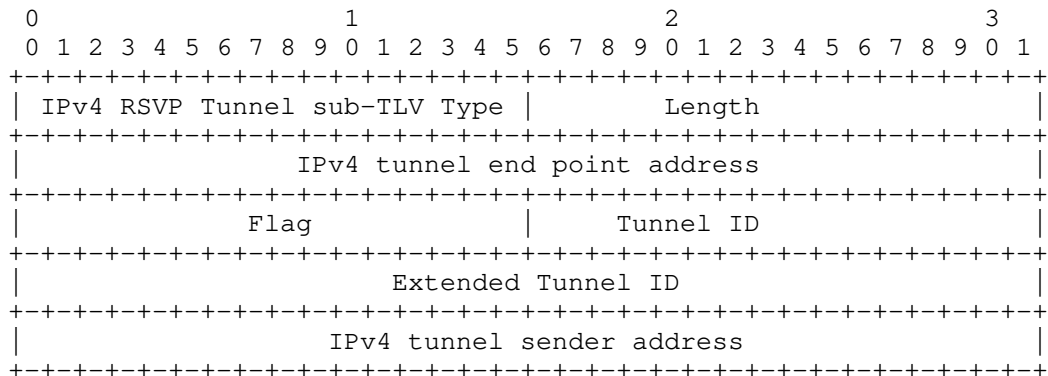
In addition, three more new sub-TLVs are defined: IPv4 RSVP Tunnel sub-TLV, IPv6 RSVP Tunnel sub-TLV, and RP TC (Traffic Class) sub-TLV. Detailed definition is in the following sections.

With the Return Path TLV flags and the sub-TLVs defined in [RFC4379] and in this document, it could provide following options for return paths specifying:

1. Specify a particular LSP as return path
  - use those sub-TLVs defined in [RFC4379],
2. Specify a more generic tunnel FEC as return path
  - use the IPv4/IPv6 RSVP Tunnel sub-TLVs defined in Section 3.3.1 and Section 3.3.2 of this document
3. Specify the reverse path of the bidirectional LSP as return path
  - use B bit defined in Section 3.2 of this document.
4. Force return path to non-default path
  - use A bit defined in Section 3.2 of this document.
5. Allow any LSPs except specific or general ones as return path
  - use E bit (Section 3.2 of this document) and combine with the specific paths identified by the sub-TLVs carried in Reply Path field.

#### 3.3.1. IPv4 RSVP Tunnel sub-TLV

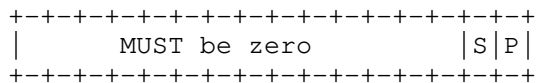
The IPv4 RSVP Tunnel sub-TLV is used in the RP TLV to allow the operator to specify a more generic tunnel FEC other than a particular LSP as the return path. The egress LSR chooses any LSP from the LSPs that have the same Tunnel attributes and satisfy the conditions carried in the Flag field. The format of IPv4 RSVP Tunnel sub-TLV is as follows:



The IPv4 RSVP Tunnel sub-TLV is derived from the RSVP IPv4 FEC TLV that is defined in Section 3.2.3 [RFC4379]. All fields have the same semantics as defined in [RFC4379] except that the LSP-ID field is omitted and a new Flag field is defined.

The IPv4 RSVP Tunnel sub-TLV Type field is 2 octets in length, and the recommended type value is 19 (to be confirmed by IANA).

The Flag field is 2 octets in length, it is used to notify the egress LSR how to choose the return path. The Flag field is a bit vector and has following format:



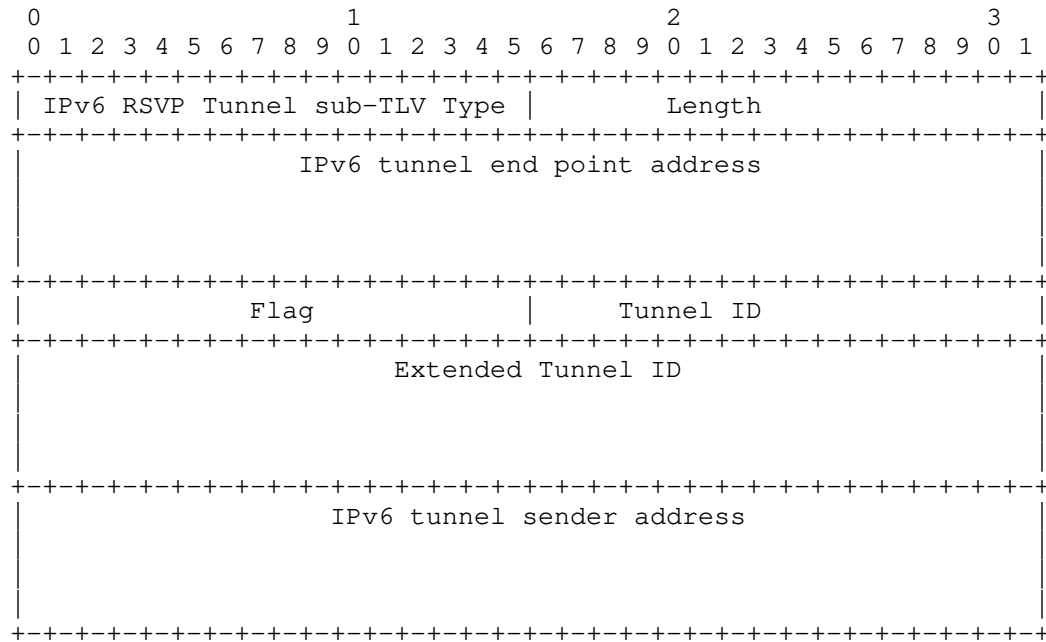
P (Primary): the return path MUST be chosen from the LSPs that have the same Tunnel attributes and the LSP MUST be the primary LSP.

S (Secondary): the return path MUST be chosen from the LSPs that have the same Tunnel attributes and the LSP MUST be the secondary LSP.

P bit and S bit MUST not both be set. If P bit and S bit are both not set, the return path could be any one of the LSPs that have the same Tunnel attributes.

## 3.3.2. IPv6 RSVP Tunnel sub-TLV

The IPv6 RSVP Tunnel sub-TLV is used in the RP TLV to allow the operator to specify a more generic tunnel FEC other than a particular LSP as the return path. The egress LSR chooses an LSP from the LSPs that have the same Tunnel attributes and satisfy the conditions carried in the Flag field. The format of IPv6 RSVP Tunnel sub-TLV is as follows:



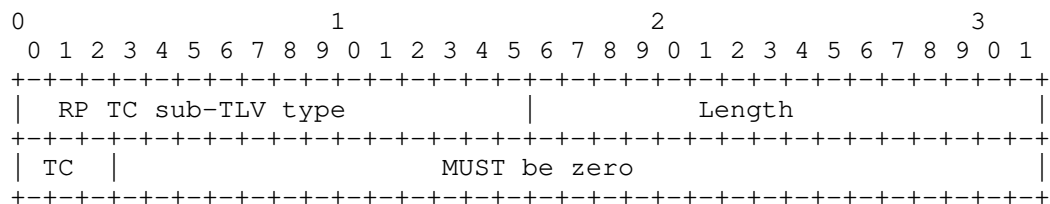
The IPv6 RSVP Tunnel sub-TLV is derived from RSVP IPv6 FEC TLV that is defined in Section 3.2.4 of [RFC4379]. All fields have the same semantics as defined in [RFC4379] except that the LSP-ID field is omitted and a new Flag field is defined..

The IPv6 RSVP Tunnel sub-TLV Type field is 2 octets in length, and the recommended type value is 20 (to be confirmed by IANA).

The Flag field is 2 octets in length and is identical to that described in Section 3.3.

### 3.3.3. RP TC sub-TLV

Reply TOS Byte TLV [RFC4379] is used by the originator of the echo request to request that an echo reply be sent with the IP header TOS byte set to the value specified in the TLV. Similarly, in this document, a new sub-TLV: RP TC sub-TLV is defined and MAY be used by the originator of the echo request to request that an echo reply be sent with the TC bits of the specified return LSP set to the value specified in this sub-TLV. Since there may be more than one FEC sub-TLVs (return paths) specified in the RP TLV, the relevant RP TC sub-TLV MUST directly follow the FEC sub-TLV that identifies the corresponding specified return LSP. The format of RP TC sub-TLV is as follows:



The RP TC sub-TLV Type field is 2 octets in length, and the recommended type value is 18 (to be confirmed by IANA).

The Length field is 2 octets in length, the value of length field is fixed 4.

## 4. Theory of Operation

The procedures defined in this document currently only apply to "ping" mode. The "traceroute" mode is out of scope for this document.

In [RFC4379], the echo reply is used to report the LSP checking result to the LSP Ping initiator. This document defines a new reply mode and a new TLV (RP TLV) which enable the LSP ping initiator to specify or constrain the return path of the echo reply. Similarly the behavior of echo reply is extended to detect the requested return path by looking at a specified path FEC TLV. This enables LSP Ping to detect failures in both directions of a path with a single operation, this of course cuts in half the operational steps required to verify the end to end bidirectional connectivity and integrity of an LSP.

When the echo reply message is intended to test the return MPLS LSP path, the destination IP address of the echo reply message MUST never be used in a forwarding decision. To avoid this possibility the destination IP address of the echo reply message that is transmitted along the specified return path MUST be set to numbers from the range 127/8 for IPv4 or 0:0:0:0:0:FFFF:127/104 for IPv6, and the IP TTL MUST be set 1. Of course when the echo reply message is not intended for testing the specified return path, the procedures defined in [RFC4379] (the destination IP address is copied from the source IP address) apply unchanged.

#### 4.1. Sending an Echo Request

When sending an echo request, in addition to the rules and procedures defined in Section 4.3 of [RFC4379], the reply mode of the echo request MUST be set to "Reply via specified path", and a RP TLV MUST be carried in the echo request message correspondingly. The RP TLV includes one or several reply path sub-TLV(s) to identify the return path(s) the egress LSR should use for its reply.

For a bidirectional LSP, since the ingress LSR and egress LSR of a bidirectional LSP are aware of the relationship between the forward and backward direction LSPs, only the B bit SHOULD be set in the RP TLV. If the operator wants the echo reply to be sent along a different path other than the reverse direction of the bidirectional LSP, "A" bit SHOULD be set or another FEC sub-TLV SHOULD be carried in the RP TLV instead, and the B bit MUST be clear.

In some cases, operators may want to treat two unidirectional LSPs (one for each direction) as a pair. There may not be any binding relationship between the two LSPs. Using the mechanism defined in this document, operators can run LSP Ping one time from one end to complete the failure detection on both unidirectional LSPs. To accomplish this, the echo request message MUST carry (in the RP TLV) a FEC sub-TLV that belongs to the backward LSP.

#### 4.2. Receiving an Echo Request

"Ping" mode processing as defined in Section 4.4 of [RFC4379] applies in this document. In addition, when an echo request is received, if the egress LSR does not know the reply mode defined in this document, an echo reply with the return code set to "Malformed echo request" and the Subcode set to zero will be send back to the ingress LSR according to the rules of [RFC4379]. If the egress LSR knows the reply mode, according to the RP TLV, it SHOULD find and select the desired return path. If there is a matched path, an echo

reply with RP TLV that identify the return path SHOULD be sent back to the ingress LSR, the RP return code SHOULD be set to "The echo reply was sent successfully using the specified RP". If there is no such path, an echo reply with RP TLV SHOULD be sent back to the ingress LSR, the RP return code SHOULD be set to relevant code (defined Section 3.2) for the real situation to reflect the result of RP TLV processing and return path selection. For example, if the specified LSP is not found, the egress then chooses another LSP as the return path to send the echo reply, the RP return code SHOULD be set to "The specified RP was not found, the echo reply was sent via other LSP", and if the egress chooses an IP path to send the echo reply, the RP return code SHOULD be set to "The specified RP was not found, the echo reply was sent via IP path". If there is unknown sub-TLV in the received RP TLV, the RP return code SHOULD be set to "One or more of the sub-TLVs in RP TLV was not understood".

If the A bit of the RP TLV in a received echo request message is set, the egress LSR SHOULD send the echo reply along an non-default return path.

If the B bit of the RP TLV in a received echo request message is set, the egress LSR SHOULD send the echo reply along the reverse direction of the bidirectional LSP.

If the E bit of the RP TLV in a received echo request message is set, the egress LSR MUST exclude the paths identified by those FEC sub-TLVs carried in the RP TLV and select other path to send the echo reply.

If the A and E bit of the RP TLV in a received echo request message is clear, the echo reply is REQUIRED not only to send along the specified path, but to test the selected return path as well (by carrying the FEC stack TLV of the return path).

In addition, the FEC validate results of the forward path LSP SHOULD not affect the egress LSR continue to test return path LSP.

#### 4.3. Sending an Echo Reply

As described in [RFC4379], the echo reply message is a UDP packet, and it MUST be sent only in response to an MPLS echo request. The source IP address is a routable IP address of the replier, the source UDP port is the well-know UDP port for LSP ping.

When the echo reply is intended to test the return path (both A and E bit are not set), the destination IP address of the echo reply

message MUST never be used in a forwarding decision. To avoid this problem, the IP destination address of the echo reply message that is transmitted along the specified return path MUST be set to numbers from the range 127/8 for IPv4 or 0:0:0:0:0:FFFF:127/104 for IPv6, and the IP TTL MUST be set 1. If the echo reply is required to test the return path, the echo reply MUST have a FEC stack TLV describing the return path, which is used for the ingress LSR to perform FEC validation. The FEC stack TLV of the forward path MUST NOT be copied to the echo reply. And the FEC stack TLV of forward LSP MUST not be copied to the echo reply.

If the echo reply message is not intended for testing the specified return path (either A bit or E bit is set), the same as defined in [RFC4379], the destination IP address and UDP port are copied from the source IP address and source UDP port of the echo request.

When sending the echo reply, a RP TLV that identifies the return path MUST be carried, the RP return code SHOULD be set to relevant code that reflects results about how the egress processes the RP TLV in a previous received echo request message and return path selection. By carrying the RP TLV in an echo reply, it gives the Ingress LSR enough information about the reverse direction of the tested path to verify the consistency of the data plane against control plane. Thus a single LSP Ping could achieve both directions of a path test.

#### 4.4. Receiving an Echo Reply

The rules and process defined in Section 4.6 of [RFC4379] apply here. When an echo reply is received, if the reply mode is "Reply via specified path" and a FEC stack TLV exists, it means that the echo reply has both Ping result reporting and reverse path checking functions. The ingress LSR MUST do FEC validation as an egress LSR does when receiving an echo request, the FEC validation process (relevant to "ping" mode) defined in Section 4.4.1 of [RFC4379] applies here.

When an echo reply is received with return code set to "Malformed echo request received" and the Subcode set to zero. It is possible that the egress LSR may not know the "Reply via specified path" reply mode, the operator may choose to re-perform another LSP Ping by using one of the four reply modes defined [RFC4379].

On receipt of an echo reply with RP return code in the RP TLV set to "The specified RP was not found, ...", it means that the egress LSR could not find a matched return path as specified. Operators may

choose to specify another LSP as the return path or use other methods to detect the path further.

When the LSP Ping initiator fails after some time to receive the echo reply message, the operator MAY initiate another LSP Ping by resending a new echo request carrying a RP TLV with E bit set, the sub-TLVs and/or B bit (when the tested LSP is a bidirectional LSP) identify the previous tried reply paths that are used to notify the egress LSR to send echo reply message along any other workable path other than these failed return paths. Hence it could improve the reliability of the echo reply message.

## 5. Security Considerations

Security considerations discussed in [RFC4379] apply to this document. In addition to that, in order to prevent using the extension defined in this document for "proxying" any possible attacks, the return path LSP MUST have destination to the same node where the forward path is from.

## 6. IANA Considerations

IANA is requested to make the following allocations from registries under its control.

### 6.1. Reply mode

IANA is requested to assign a new reply mode as follows:

Reply mode:

Value	Meaning
-----	-----
5	Reply via specified path

### 6.2. RP TLV

IANA is requested to assign a new TLV type (TBD) from the range of 0-16383. We suggest that the value 20 be assigned for the new RP TLV type.

Type	Value Field
-----	-----
20	Reply Path

### 6.3. Sub-TLVs for RP TLV

This document defines four new sub-TLV Types (described in Section 3.4, 3.5, 3.6 and 3.7) of RP TLV, and those FEC sub-TLVs defined in [RFC4379] are applicable for inclusion in RP TVL.

IANA is requested to assign sub-TLVs as follows. The following numbers are suggested:

Sub-type	Value Field	Reference
-----	-----	-----
17	RP TC	this document
18	IPv4 RSVP Tunnel	this document
19	IPv6 RSVP Tunnel	this document

### 7. Contributors

The following individuals also contributed to this document:

Ehud Doron  
Orckit-Corrigent

EMail: ehudd@orckit.com

Ronen Solomon  
Orckit-Corrigent

EMail: RonenS@orckit.com

Ville Hallivuori  
Tellabs  
Sinimaentie 6 C  
FI-02630 Espoo, Finland

EMail: ville.hallivuori@tellabs.com

## 8. Acknowledgments

The authors would like to thank Adrian Farrel and Peter Ashwood-Smith for their review, suggestion and comments to this document.

## 9. References

### 9.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC4379] K. Kompella., et al., "Detecting Multi-Protocol Label Switched (MPLS) Data Plane Failures", RFC 4379, February 2006.
- [BFD] D. Katz, D. and Ward, D., "Bidirectional Forwarding Detection", draft-ietf-bfd-base, work in progress.
- [BFD-MPLS] Aggarwal, R., Kompella, K., Nadeau, T., and Swallow, G., "BFD For MPLS LSPs", draft-ietf-bfd-mpls, work in progress.
- [BFD-IP] D. Katz, D. Ward, "BFD for IPv4 and IPv6 (Single Hop)", draft-ietf-bfd-v4v6-lhop-08.txt.

## 9.2. Informative References

- [RFC3471] L. Berger, "Generalized Multi-Protocol Label Switching (GMPLS) Signaling Functional Description", RFC 3471, January 2003.
- [RFC3473] L. Berger, "Generalized Multi-Protocol Label Switching (GMPLS) Signaling", RFC 3473, January 2003.
- [RFC3945] E. Mannie, "Generalized Multi-Protocol Label Switching (GMPLS) Architecture", RFC 3945, October 2004.
- [RFC5654] Niven-Jenkins, B. (Ed.), Brungard, D. (Ed.), Betts, M. (Ed.) Sprecher, N., and Ueno, S., "Requirements of an MPLS Transport Profile", RFC 5654, September 2009.

## Authors' Addresses

Mach(Guoyi) Chen  
Huawei Technologies Co., Ltd.  
No. 3 Xinxu Road  
Shangdi Information Industry Base  
Hai-Dian District, Beijing 100085  
China

EMail: mach@huawei.com

So Ning  
Verizon Inc.  
2400 N. Glenville Rd.,  
Richardson, TX 75082

Phone: +1 972-729-7905  
EMail: ning.so@verizonbusiness.com

Frederic Jounay  
France Telecom  
2, avenue Pierre-Marzin  
22307 Lannion Cedex  
FRANCE

EMail: frederic.jounay@orange-ftgroup.com

Simon Delord  
Telstra  
242 Exhibition St  
Melbourne VIC 3000  
Australia

EMail: simon.a.delord@team.telstra.com

Xinchun Guo  
Huawei Technologies Co., Ltd.  
No. 3 Xinxu Road  
Shangdi Information Industry Base  
Hai-Dian District, Beijing 100085  
China

EMail: guoxinchun@huawei.com

Wei Cao  
Huawei Technologies Co., Ltd.  
No. 3 Xixi Road  
Shangdi Information Industry Base  
Hai-Dian District, Beijing 100085  
China

EMail: caoweigne@huawei.com



MPLS Working Group  
Internet Draft  
Intended status: Standards Track  
Expires: April 2011

Dave Allan, Ed.  
Ericsson  
  
George Swallow Ed.  
Cisco Systems, Inc

John Drake Ed.  
Juniper

October 22, 2010

Proactive Connectivity Verification, Continuity Check and Remote  
Defect indication for MPLS Transport Profile  
draft-ietf-mpls-tp-cc-cv-rdi-02

Abstract

Continuity Check (CC), Proactive Connectivity Verification (CV) and Remote Defect Indication (RDI) functionalities are required for MPLS-TP OAM.

Continuity Check monitors the integrity of the continuity of the LSP for any loss of continuity defect. Connectivity verification monitors the integrity of the routing of the LSP between sink and source for any connectivity issues. RDI enables an End Point to report, to its associated End Point, a fault or defect condition that it detects on a PW, LSP or Section.

This document specifies methods for proactive CV, CC, and RDI for MPLS-TP Label Switched Path (LSP), PWs and Sections using Bidirectional Forwarding Detection (BFD).

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 RFC2119 [1].

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 November 28, 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
1.1. Authors.....	4
2. Conventions used in this document.....	4
2.1. Terminology.....	4
2.2. Issues for discussion.....	5
3. MPLS CC, proactive CV and RDI Mechanism using BFD.....	5
3.1. ACH code points for CC and proactive CV.....	6
3.2. MPLS BFD CC Message format.....	6
3.3. MPLS BFD proactive CV Message format.....	7
3.4. BFD Session in MPLS-TP terminology.....	7
3.5. BFD Profile for MPLS-TP.....	8
3.5.1. Session initiation.....	9
3.5.2. Defect entry criteria.....	9

3.5.3. Defect entry consequent action.....	10
3.5.4. Defect exit criteria.....	11
3.5.5. State machines.....	11
3.5.6. Configuration of MPLS-TP BFD sessions.....	14
3.5.7. Discriminator values.....	14
4. Acknowledgments.....	15
5. IANA Considerations.....	15
6. Security Considerations.....	15
7. References.....	15
7.1. Normative References.....	15
7.2. Informative References.....	16

## 1. Introduction

In traditional transport networks, circuits are provisioned on two or more switches. Service Providers (SP) need OAM tools to detect mis-connectivity and loss of continuity of transport circuits. Both PWs and MPLS-TP LSPs [7] emulating traditional transport circuits need to provide the same CC and proactive CV capabilities as required in draft-ietf-mpls-tp-oam-requirements[3]. This document describes the use of BFD for CC, proactive CV, and RDI of a PW, LSP or SPME between two Maintenance Entity Group End Points (MEPs).

As described in [9], Continuity Check (CC) and Proactive Connectivity Verification (CV) functions are used to detect loss of continuity (LOC), and unintended connectivity between two MEPs (e.g. mismerging or misconnectivity or unexpected MEP).

The Remote Defect Indication (RDI) is an indicator that is transmitted by a MEP to communicate to its peer MEP that a signal fail condition exists. RDI is only used for bidirectional LSPs and is associated with proactive CC & CV packet generation.

This document specifies the BFD extension and behavior to satisfy the CC, proactive CV monitoring and the RDI functional requirements for both co-routed and associated bi-directional LSPs. Supported encapsulations include GAL/G-ACh, VCCV and UDP/IP. Procedures for uni-directional LSPs are for further study.

The mechanisms specified in this document are restricted to BFD asynchronous mode.

### 1.1. Authors

David Allan, John Drake, George Swallow, Annamaria Fulignoli, Sami Boutros, Siva Sivabalan, David Ward, Martin Vigoureux.

## 2. Conventions used in this document

### 2.1. Terminology

ACH: Associated Channel Header

BFD: Bidirectional Forwarding Detection

CV: Connectivity Verification

GAL: Generalized Alert Label

LDI: Link Down Indication

LKI: Lock Instruct

LKR: Lock Report

LSR: Label Switching Router

MEG: Maintenance Entity Group

MEP: Maintenance Entity Group End Point

MIP: Maintenance Entity Group Intermediate Point

MPLS-OAM: MPLS Operations, Administration and Maintenance

MPLS-TP: MPLS Transport Profile

MPLS-TP LSP: Uni-directional or Bidirectional Label Switch Path representing a circuit

MS-PW: Multi-Segment PseudoWire

NMS: Network Management System

PW: Pseudo Wire

RDI: Remote Defect Indication.

SPME: Sub-Path Maintenance Entity

TTL: Time To Live

TLV: Type Length Value

VCCV: Virtual Circuit Connectivity Verification

## 2.2. Issues for discussion

### 1) Requirement for additional BFD diagnostic codes?

#### 1. When periodicity of CV cannot be supported

## 3. MPLS CC, proactive CV and RDI Mechanism using BFD

This document proposes distinct encapsulations and code points for ACh encapsulated BFD depending on whether the mode of operation is CC or CV:

- o CC mode: defines a new code point in the Associated Channel Header (ACH) described in [2]. In this mode Continuity Check and RDI functionalities are supported.
- o CV mode: defines a new code point in the Associated Channel Header (ACH) described in [2]. The ACH with "MPLS Proactive CV" code point indicates that the message is an MPLS BFD proactive CV and CC message and CC, CV and RDI functionalities are supported.

RDI: is communicated via the BFD diagnostic field in BFD CC and CV messages. It is not a distinct PDU. A sink MEP will encode a diagnostic code of "1- Control detection time expired" when the interval times detect multiplier have been exceeded, and with "3 - neighbor signaled session down" as a consequence of the sink MEP receiving AIS with LDI set. A sink MEP that has started sending diag code 3 will NOT change it to 1 when the detection timer expires.

In accordance with RFC 5586, when these packets are encapsulated in an IP header the fields in the IP header are set as defined in RFC 5884. It should also be noted that existing ACh code points and mechanisms for negotiating the control channel and connectivity verification (i.e. OAM functions) between PEs are specified for VCCV[6].

### 3.1. ACH code points for CC and proactive CV

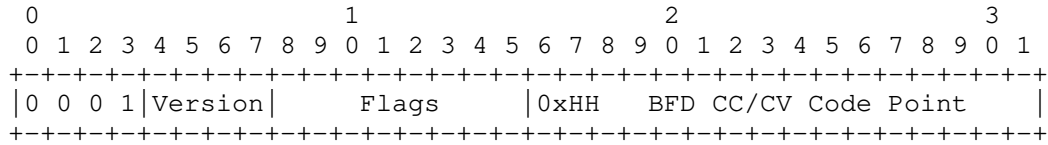


Figure 1: ACH Indication of MPLS-TP Connectivity Verification

The first nibble (0001b) indicates the ACH.

The version and the flags are set to 0 as specified in [2].

The code point is either

- BFD CC code point = 0xHH. [HH to be assigned by IANA from the PW Associated Channel Type registry.] or,
- BFD proactive CV code point = 0xHH. [HH to be assigned by IANA from the PW Associated Channel Type registry.]

Both CC and CV modes apply to PWs, MPLS LSPs (including tandem connection monitoring), and Sections.

CC and CV operation can be simultaneously employed on an ME within a single BFD session. The expected usage is that normal operation is to send CC BFD PDUs with every nth BFD PDU augmented with a source MEP ID and identified as requiring additional processing by the different ACh channel type.

### 3.2. MPLS BFD CC Message format

The format of an MPLS CC Message is shown below.

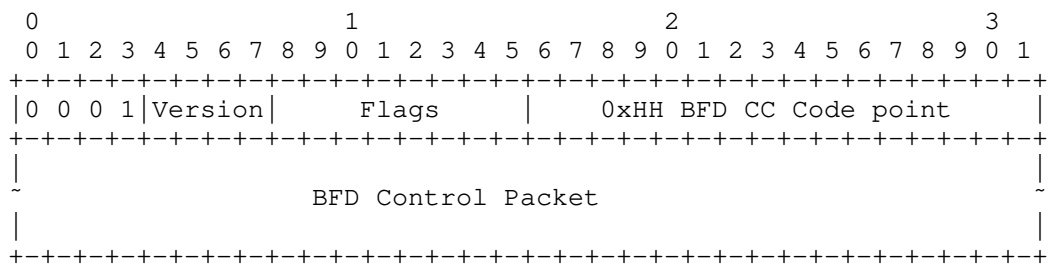


Figure 2: MPLS CC Message

### 3.3. MPLS BFD proactive CV Message format

The format of an MPLS CV Message is shown below.

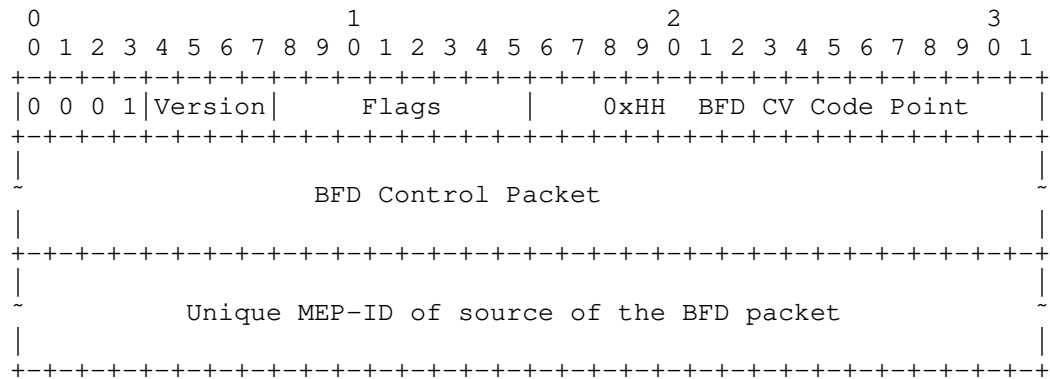


Figure 3: MPLS CV Message

As shown in Figure 3, BFD Control packet as defined in [4] is transmitted as MPLS labeled packets along with the ACH. Appended to the BFD control packet is a MEP Source ID TLV. The length in the BFD control packet is as per [4]. There are 4 possible Source MEP TLVs (corresponding to the MEP IDs defined in [8] [type fields to be assigned by IANA]. The type fields are:

- X1 - ICC encoded MEP ID
- X2 - LSP-MEP\_ID
- X3 - PW MEP ID
- X4 - PW Segment endpoint ID

When GAL label is used, the TTL field of the GAL MUST be set to at least 1, and the GAL will be the end of stack label (S=1).

### 3.4. BFD Session in MPLS-TP terminology

A BFD session corresponds to a CC or a proactive CV OAM instance in MPLS-TP terminology.

A BFD session is enabled when the CC or proactive CV functionality is enabled on a configured Maintenance Entity (ME)..

On a Sink MEP, a BFD session can be in DOWN, INIT or UP state as detailed in [4].

When on a ME the CC or proactive CV functionality is disabled, the BFD session transitions to the ADMIN DOWN State and the BFD session ends.

A new BFD session is initiated when the operator enables or re-enables the CC or CV functionality on the same ME.

### 3.5. BFD Profile for MPLS-TP

BFD MUST operate in asynchronous mode. In this mode, the BFD Control packets are periodically sent at configurable time rate. This rate is typically a fixed value for the lifetime of the session. In the rare circumstance where an operator has a reason to change session parameters, the session must be moved to the ADMIN DOWN state. Poll/final discipline can only be used for VCCV and UDP/IP encapsulated BFD.

The transport profile is designed to operate independent of the control plane; hence the C bit SHOULD be set.

This document specifies bi-directional BFD for p2p transport LSPs, hence the M bit MUST be clear.

There are two modes of operation for bi-directional LSPs. One in which the session state of both directions of the LSP is coordinated and one constructed from BFD sessions in such a way that the two directions operate independently. A single bi-directional BFD session is used for coordinated operation. Two independent BFD sessions are used for independent operation.

Coordinated operation is as described in [4]. Independent operation requires clarification of two aspects of [4]. Independent operation is characterized by the setting of MinRxInterval to zero by the MEP that is typically the session originator (referred to as the source MEP), and there will be a session originator at either end of the bi-directional LSP. Each source MEP will have a corresponding sink MEP that has been configured to a Tx interval of zero.

The base spec is unclear on aspects of how a MEP with a BFD transmit rate set to zero behaves. One interpretation is that no periodic messages on the reverse component of the bi-directional LSP originate with that MEP, it will only originate messages on a state change.

The first clarification is that when a state change occurs a MEP set to a transmit rate of zero sends BFD control messages with a one

second period on the reverse component until such time that the state change is confirmed by the session peer. At this point the MEP set to a transmit rate of zero can resume quiescent behavior. This adds robustness to all state transitions in the RxInterval=0 case.

The second is that the originating MEP (the one with a non-zero TxInterval) will ignore a DOWN state received from a zero interval peer. This means that the zero interval peer will continue to send DOWN state messages that include the RDI diagnostic code as the state change is never confirmed. This adds robustness to the exchange of RDI indication on a uni-directional failure (for both session types DOWN with a diagnostic of either control detection period expired or neighbor signaled session down offering RDI functionality).

A further extension to the base specification is that there are additional OAM protocol exchanges that act as inputs to the BFD state machine; these are the Link Down Indication [5] and the Lock Instruct/Lock Report transactions; Lock Report interaction being optional.

### 3.5.1. Session initiation

In all scenarios a BFD session starts with both ends in the DOWN state. DOWN state messages exchanged include the desired Tx and Rx rates for the session. If a node cannot support the Min Tx rate desired by a peer MEP it does not transition from down to the INIT state and sends a diagnostic code (TBD) indicating that the requested Tx rate cannot be supported.

Otherwise once a transition from DOWN to INIT has occurred, the session progresses as per [4]. In both the DOWN and INIT states messages are transmitted at a rate of one per second and the defect detection interval is fixed at 3.5 seconds. On transition to the UP state message periodicity changes to the negotiated rate and the detect interval switches to detect multiplier times the session peer's Tx Rate.

### 3.5.2. Defect entry criteria

There are further defect criteria beyond those that are defined in [4] to consider given the possibility of mis-connectivity and mis-configuration defects. The result is the criteria for a LSP direction to transition from the defect free state to a defect state is a superset of that in the BFD base specification [4]. The following conditions cause a MEP to enter the defect state for CC or CV:

1. BFD session times out (Loss of Continuity defect).

2. Receipt of a link down indication.
3. Receipt of an unexpected M bit (Session Mis-configuration defect).

And the following will cause the MEP to enter the defect state for CV operation

1. BFD control packets are received with an unexpected encapsulation (mis-connectivity defect), these include:
  - a PW receiving a packet with a GAL
  - an LSP receiving an IP header instead of a GAL (note there are other possibilities that can also alias as an OAM packet)
2. Receipt of an unexpected globally unique Source MEP identifier (Mis-connectivity defect).
3. Receipt of an unexpected session discriminator in the your discriminator field (mis-connectivity defect).
4. Receipt of an expected session discriminator with an unexpected label (mis-connectivity defect).

The effective defect hierarchy (order of checking) is

1. Receiving nothing.
2. Receiving link down indication.
3. Receiving from an incorrect source (determined by whatever means).
4. Receiving from a correct source (as near as can be determined), but with incorrect session information).
5. Receiving control packets in all discernable ways correct.

### 3.5.3. Defect entry consequent action

Upon defect entry a sink MEP will assert signal fail into any client (sub-)layers. It will also communicate session DOWN to its session peer.

The blocking of traffic as consequent action MUST be driven only by a defect's consequent action as specified in draft-ietf-mpls-tp-oam-framework [9] section 5.1.1.2.

When the defect is mis-branching, the LSP termination will silently discard all non-oam traffic received.

#### 3.5.4. Defect exit criteria

##### 3.5.4.1. Exit from a Loss of continuity defect

For a coordinated session, exit from a loss of connectivity defect is as described in figure 4 which updates [4].

For an independent session, exit from a loss of connectivity defect occurs upon receipt of a well formed control packet from the peer MEP as described in figures 5 and 6.

##### 3.5.4.2. Exit from a session mis-configuration defect

[editors: for a future version of the document]

##### 3.5.4.3. Exit from a mis-connectivity defect

[Editors note: The shift to CC with interleaved CV suggests the CV periodicity may not be known by a sink MEP, hence exit criteria from a mis-connectivity defect may not be able to be established. We suggest two possible resolutions for this:

1. Exit criteria is manual intervention.
2. A minimum CV insertion rate (say 1/sec) be specified such that the exit criteria be specified as no mis-connected CV PDUs be received for a minimum of 3 times the minimum insertion rate]

#### 3.5.5. State machines

The following state machines update [4]. They have been modified to include AIS with LDI set and LKI as inputs to the state machine and to clarify the behavior for independent mode. LKR is an optional input.

The coordinated session state machine has been augmented to indicate AIS with LDI set and optionally LKR as inputs to the state machine. For a session that is in the UP state, receipt of AIS with LDI set or optionally LKR will transition the session into the DOWN state.

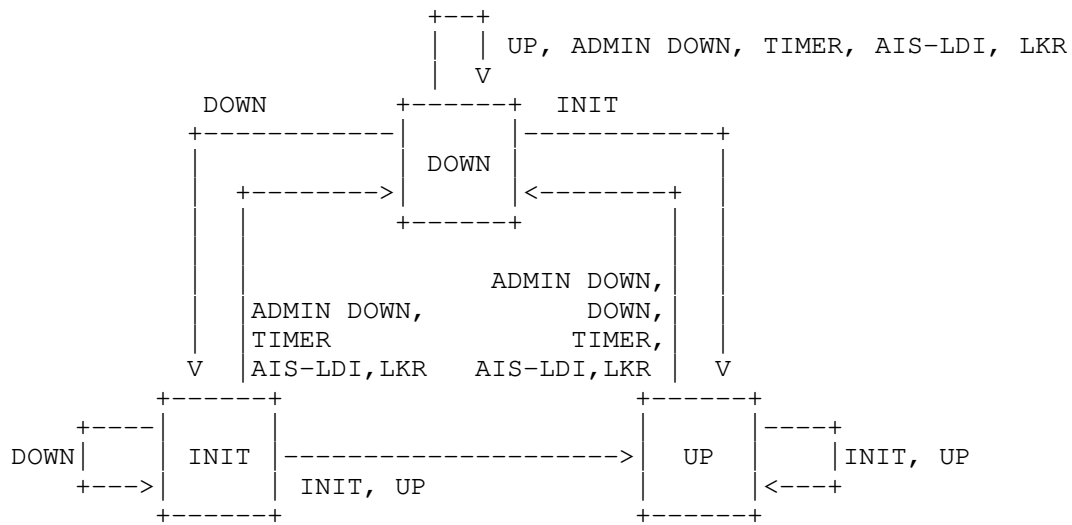


Figure 4: State machine for coordinated session operation

For independent mode, there are two state machines. One for the source MEP (who requested `MinRxInterval=0`) and the sink MEP (who agreed to `MinRxInterval=0`).

The source MEP will not transition out of the UP state once initialized except in the case of a forced ADMIN DOWN. Hence AIS-with LDI set and optionally LKR do not enter into the state machine transition from the UP state, but do enter into the INIT and DOWN states.

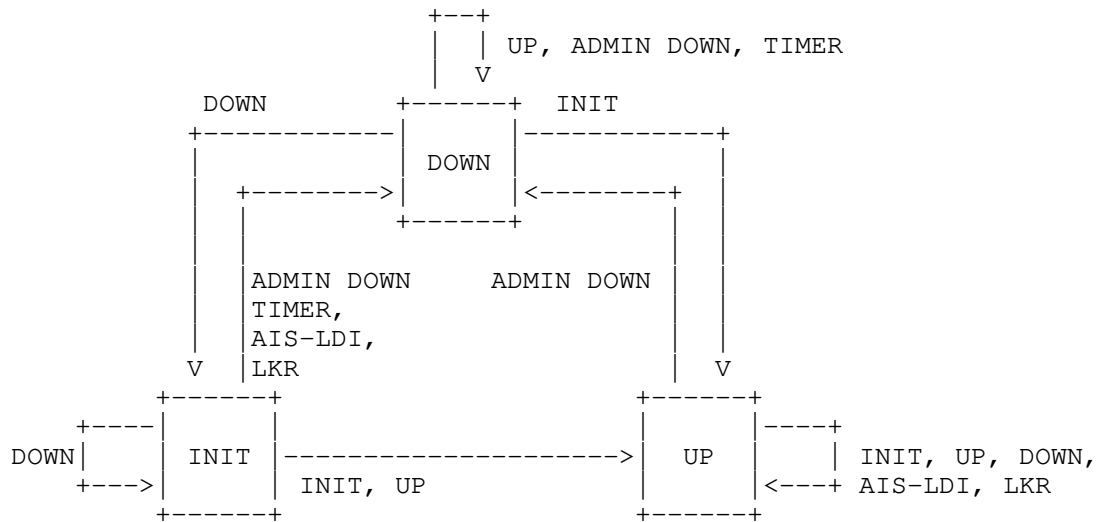


Figure 5: State machine for source MEP for independent session operation

The sink MEP state machine (for which the transmit interval has been set to zero) is modified to:

- 1) Permit direct transition from DOWN to UP once the session has been initialized. With the exception of via the ADMIN DOWN state, the source MEP will never transition from the UP state, hence in normal unidirectional fault scenarios will never transition to the INIT state.

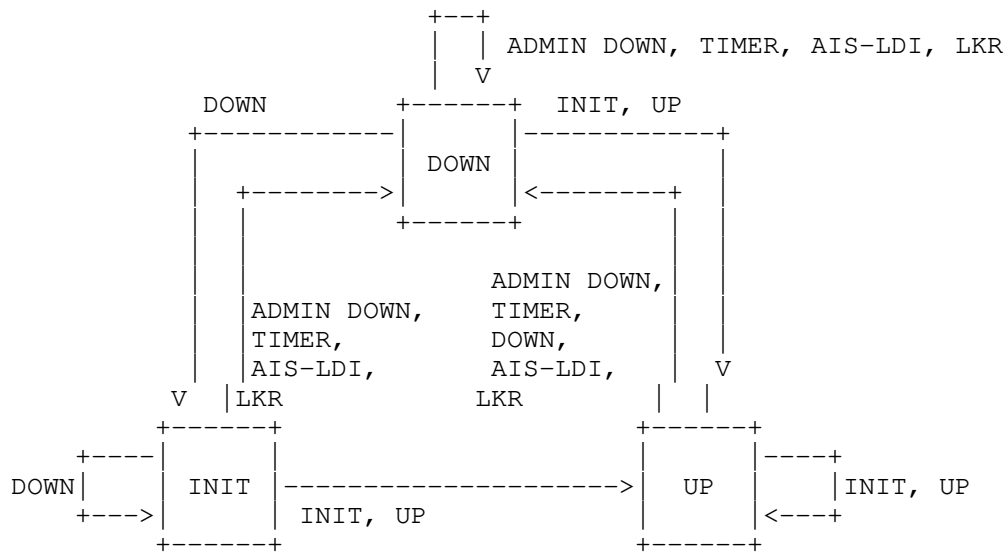


Figure 6: State machine for the sink MEP for independent session operation

### 3.5.6. Configuration of MPLS-TP BFD sessions

[Editors note, for a future revision of the document]

### 3.5.7. Discriminator values

In the BFD control packet the discriminator values have either local to the sink MEP or no significance (when not known).

My Discriminator field MUST be set to a nonzero value (it can be a fixed value), the transmitted your discriminator value MUST reflect back the received value of My discriminator field or be set to 0 if that value is not known.

Although the BFD base specification permits an implementation to change the my discriminator field at arbitrary times, this is not permitted for CV mode in order to avoid race conditions in mis-connectivity defects.

#### 4. Acknowledgments

To be added in a later version of this document

#### 5. IANA Considerations

To be added in a later version of this document

#### 6. Security Considerations

The security considerations for the authentication TLV need further study.

Base BFD foresees an optional authentication section (see [4] section 6.7); that can be extended also to the tool proposed in this document.

Authentication methods that require checksum calculation on the outgoing packet must extend the checksum also on the ME Identifier Section. This is possible but seems uncorrelated with the solution proposed in this document: it could be better to use the simple password authentication method.

#### 7. References

##### 7.1. Normative References

- [1] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [2] Bocci, M. et al., " MPLS Generic Associated Channel ", RFC 5586 , June 2009
- [3] Vigoureux, M., Betts, M. and D. Ward, "Requirements for Operations Administration and Maintenance in MPLS Transport Networks", RFC5860, May 2010
- [4] Katz, D. and D. Ward, "Bidirectional Forwarding Detection", RFC 5880, June 2010
- [5] Swallow, G. et al., "MPLS Fault Management OAM", draft-ietf-mpls-tp-fault-02 (work in progress), July 2010
- [6] Nadeau, T. and C. Pignataro, "Pseudowire Virtual Circuit Connectivity Verification (VCCV): A Control Channel for Pseudowires", RFC 5085, December 2007

## 7.2. Informative References

- [7] Bocci, M., et al., "A Framework for MPLS in Transport Networks", RFC5921, July 2010
- [8] Bocci, M. and G. Swallow, "MPLS-TP Identifiers", draft-swallow-mpls-tp-identifiers-02 (work in progress), July 2010
- [9] Allan, D., and Busi, I. "MPLS-TP OAM Framework", draft-ietf-mpls-tp-oam-framework-09 (work in progress), October 2010

## Authors' Addresses

Dave Allan  
Ericsson  
Email: david.i.allan@ericsson.com

John Drake  
Juniper  
Email: jdrake@juniper.net

George Swallow  
Cisco Systems, Inc.  
Email: swallow@cisco.com

Annamaria Fulignoli  
Ericsson  
Email: annamaria.fulignoli@ericsson.com

Sami Boutros  
Cisco Systems, Inc.  
Email: sboutros@cisco.com

Martin Vigoureux  
Alcatel-Lucent  
Email: martin.vigoureux@alcatel-lucent.com

Siva Sivabalan  
Cisco Systems, Inc.  
Email: msiva@cisco.com

David Ward  
Juniper  
Email: dward@juniper.net



MPLS Working Group  
Internet-Draft  
Intended status: Standards Track  
Expires: April 28, 2011

G. Swallow, Ed.  
Cisco Systems, Inc.  
A. Fulignoli, Ed.  
Ericsson  
M. Vigoureux, Ed.  
Alcatel-Lucent  
S. Boutros  
Cisco Systems, Inc.  
D. Ward  
Juniper Networks, Inc.

October 25, 2010

MPLS Fault Management OAM  
draft-ietf-mpls-tp-fault-03

Abstract

This draft specifies OAM messages to indicate service disruptive conditions for MPLS Transport Profile (MPLS-TP) Label Switched Paths (LSPs). The notification mechanism employs a generic method for a service disruptive condition to be communicated to a Maintenance End Point (MEP). An MPLS Operation, Administration, and Maintenance (OAM) channel is defined along with messages to communicate various types of service disruptive conditions.

Status of this Memo

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

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

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

This Internet-Draft will expire on April 28, 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.

## Table of Contents

1. Introduction . . . . .	3
1.1. Terminology . . . . .	3
2. MPLS Fault Management Messages . . . . .	4
2.1. MPLS-TP Alarm Indication Signal . . . . .	5
2.1.1. MPLS-TP Link Down Indication . . . . .	5
2.2. MPLS-TP Lock Report . . . . .	6
3. MPLS Fault Management Channel . . . . .	6
4. MPLS Fault Management Message Format . . . . .	7
4.1. Fault Management Message TLVs . . . . .	8
4.1.1. Interface Identifier TLV . . . . .	9
4.1.2. Global Identifier . . . . .	10
4.1.3. International Carrier Code . . . . .	10
5. Sending and Receiving Fault Management Messages . . . . .	10
5.1. Sending a Fault Management Message . . . . .	10
5.2. Clearing a FM Indication . . . . .	11
5.3. Receiving a FM Indication . . . . .	11
6. Minimum Implementation Requirements . . . . .	11
7. Security Considerations . . . . .	12
8. IANA Considerations . . . . .	12
8.1. Pseudowire Associated Channel Type . . . . .	12
8.2. MPLS Fault OAM Message Type Registry . . . . .	12
8.3. MPLS Fault OAM TLV Registry . . . . .	13
9. References . . . . .	13
9.1. Normative References . . . . .	13
9.2. Informative References . . . . .	14
Authors' Addresses . . . . .	14

## 1. Introduction

In traditional transport networks, circuits such as T1 lines are provisioned on multiple switches. When a disruption occurs on any link or node along the path of such a transport circuit, OAM are generated which may in turn suppress alarms and/or activate a backup circuit. The MPLS Transport Profile (MPLS-TP) provides mechanisms to emulate traditional transport circuits. Therefore a Fault Management (FM) capability must be defined for MPLS. This capability is being defined to meet the MPLS-TP requirements as defined in RFC 5654 [1], and the MPLS-TP Operations, Administration and Maintenance Requirements as defined in RFC 5860 [2]. However, this mechanism is intended to be applicable to other aspects of MPLS as well.

Two broad classes of service disruptive conditions are identified.

1. Defect: the situation in which the density of anomalies has reached a level where the ability to perform a required function has been interrupted.
2. Lock: an administrative status in which it is expected that only test traffic, if any, and OAM (dedicated to the LSP) can be sent on an LSP.

Within the Defect class, a further category, Fault is identified. A fault is the inability of a function to perform a required action. A fault is a persistent defect.

This document specifies an MPLS OAM channel called an "MPLS-OAM Fault Management (FM)" channel. A single message format and a set of procedures are defined to communicate service disruptive conditions from the location where they occur to the endpoints of LSPs which are affected by those conditions. Multiple message types and flags are used to indicate and qualify the particular condition.

Corresponding to the two classes of service disruptive conditions listed above, two messages are defined to communicate the type of condition. These are known as:

Alarm Indication Signal (AIS)

Lock Report (LKR)

### 1.1. Terminology

ACH: Associated Channel Header

ASN: Autonomous System Number

CC: Continuity Check

FM: Fault Management

GAL: Generic Associated Channel Label

LOC: Loss of Continuity

LSP: Label Switched Path

LSR: Label Switching Router

MEP: Maintenance End Point

MIP: Maintenance Intermediate Point

MPLS: Multi-Protocol Label Switching

MPLS-TP: MPLS Transport Profile

OAM: Operations, Administration and Maintenance

P2MP: Point to Multi-Point

P2P: Point to Point

PSC: Protection State Coordination

PW: Pseudowire

TLV: Type Length Value

TTL: Time To Live

#### 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 [3].

## 2. MPLS Fault Management Messages

This document defines messages to indicate service disruptive conditions. Two messages are defined, Alarm Indication Signal, and Lock Report. These semantics of the individual messages are described in subsections below.

Fault Management messages are carried in-band by using the Associated Channel Header (ACH) and Generic Associated Channel Label (GAL) as defined in RFC5586 [4]. To facilitate recognition and delivery of Fault Management messages, the Fault Management Channel is identified by a unique ACH codepoint.

Fault OAM messages are generated by intermediate nodes where an LSP is switched. When a server layer, (e.g. a link) used by the LSP fails, the intermediate node sends Fault Management messages using the LSP's Fault associated channel back to the endpoint of the LSP. Strictly speaking, when a server MEP detects a service disruptive condition, Fault Management messages are generated by the convergence server-to-client adaptation function. The messages are sent to the client MEPs by inserting them into the affected LSPs in the direction opposite to the detecting MEP's peer server MEP(s). The message is sent periodically until the condition is cleared.

## 2.1. MPLS-TP Alarm Indication Signal

The MPLS-TP Alarm Indication Signal (AIS) message is generated in response to detecting defects in the server layer. The AIS message SHOULD be sent as soon as the condition is detected, that is before any determination has been made as to whether the condition is persistent and therefore fatal. For example, an AIS message may be sent during a protection switching event and would cease being sent (or cease being forwarded by the protection switch selector) if the protection switch was successful in restoring the link.

The primary purpose of the AIS message is to suppress alarms in the MPLS-TP layer network above the level at which the defect occurs. When the Link Down Indication is set, the AIS message MAY be used to trigger recovery mechanisms.

### 2.1.1. MPLS-TP Link Down Indication

The LDI flag is set in response to detecting a fatal failure in the server layer. The LDI flag MUST NOT be set until the defect has been determined to be fatal. The LDI flag MUST be set if the defect has been determined to be fatal. For example during a protection switching event the LDI flag is not set. However if the protection switch was unsuccessful in restoring the link within the expected repair time, the LDI flag MUST be set.

The setting of the LDI flag can be predetermined based on the protection state. If the Server Layer is protected and both the working and protection paths are available, both the active and standby MEPs should be programmed to send AIS with LDI clear upon detecting a defect condition. If the Server Layer is unprotected or

the Server Layer is protected but only the active path is available, the active MEP should be programmed to send AIS with LDI set upon detecting a defect condition.

The receipt of an AIS message with the LDI flag set MAY be treated as the equivalent of loss of continuity (LOC) at the client layer. The choice of treatment is related to the rate at which the Continuity Check (CC) function is running. In a normal transport environment, CC is run at a high rate in order to detect a failure within 10s of milliseconds. In such an environment, the LDI flag may be ignored. AIS messages with the LDI flag set SHOULD be treated the same as any other AIS message, that is, used solely for alarm suppression.

In more general MPLS environments the CC function may be running at a much slower rate. In this environment, the LDI flag enables faster switch-over upon a failure occurring along the LSP.

## 2.2. MPLS-TP Lock Report

The MPLS-TP Lock Report (LKR) message is generated when a server layer entity has been administratively locked to communicate that condition to inform the client layer entities of that condition. When an MPLS-TP LSP is administratively locked it is not available to carry client traffic. The purpose of the LKR message is to suppress alarms in the MPLS-TP layer network above the level at which the defect occurs and to allow the clients to differentiate the lock condition from a defect condition.

The primary purpose of the LKR message is to suppress alarms in the MPLS-TP layer network above the level at which the defect occurs. Like AIS with the LDI flag set, the receipt of an LKR message MAY be treated as the equivalent of loss of continuity at the client layer.

## 3. MPLS Fault Management Channel

The MPLS Fault Management channel is identified by the ACH as defined in RFC 5586 [4] with the Channel Type set to the MPLS Fault Management (FM) code point = 0xHH. [HH to be assigned by IANA from the PW Associated Channel Type registry. Note: An early codepoint allocation has made: 0x0058 Fault OAM (TEMPORARY - expires 2011-07-16)] The FM Channel does not use ACH TLVs and MUST not include the ACH TLV header. The FM ACH Channel is shown below.

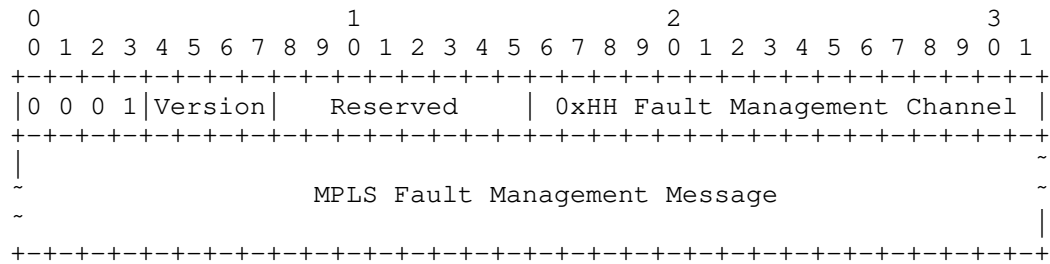


Figure 1: ACH Indication of the MPLS-TP Fault Management Channel

The Fault Management Channel is 0xHH (to be assigned by IANA)

#### 4. MPLS Fault Management Message Format

The format of the Fault Management message is shown below.

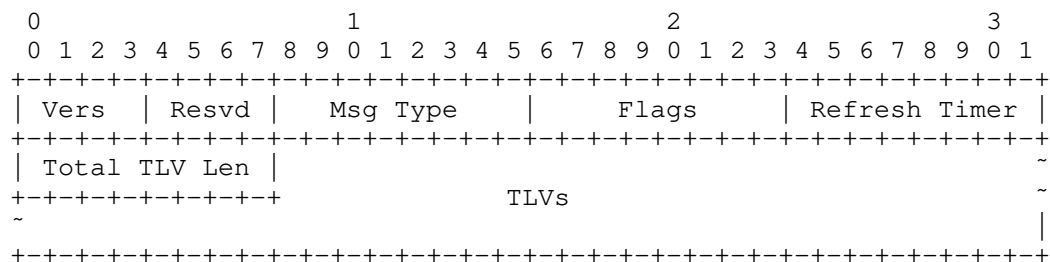


Figure 2: MPLS-TP OAM Message Format

##### Version

The Version Number is currently 1.

##### Reserved

This field MUST be set to zero on transmission and ignored on receipt.

##### Message Type

The Message Type indicates the type of condition as listed in the table below.

Msg Type	Description
-----	-----
0x0	Reserved
0x1	Alarm Indication Signal (AIS)
0x2	Lock Report (LKR)

#### Refresh Timer

The maximum time between successive FM messages specified in seconds. The range is 1 to 20. The value 0 is not permitted.

#### Total TLV Length

The total TLV length is the total of all included TLVs.

#### Flags

Two flags are defined. The reserved flags in this field MUST be set to zero on transmission and ignored on receipt.

```

+---+---+---+---+---+
| Reserved |L|R|
+---+---+---+---+---+

```

Figure 3: Flags

#### L-flag

Link Down Indication. See section Section 2.1.1 for details on setting this bit.

#### R-flag

The R-flag is normally set to zero. A setting of one indicates the removal of a previously sent FM condition.

### 4.1. Fault Management Message TLVs

TLVs are used in fault OAM to carry information that may not pertain to all messages as well as to allow for extensibility. The TLVs currently defined are the IF\_ID, Global-ID, and ICC.

TLVs (Type-Length-Value tuples) have the following format:

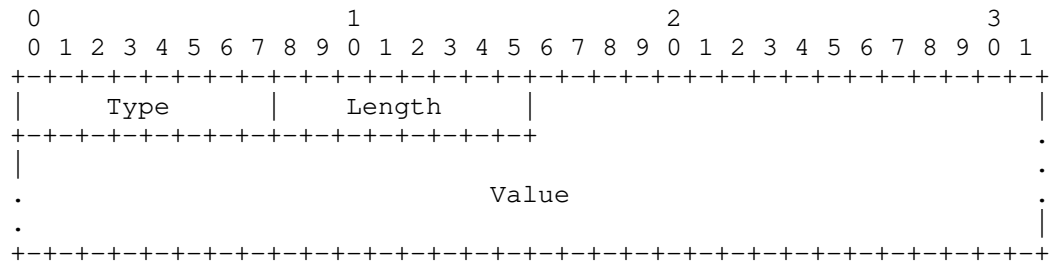


Figure 4: Fault TLV Format

**Type**

Encodes how the Value field is to be interpreted.

**Length**

Specifies the length of the Value field in octets.

**Value**

Octet string of Length octets that encodes information to be interpreted as specified by the Type field.

**4.1.1.1. Interface Identifier TLV**

This TLV carries the Interface Identifier as defined in draft-ietf-mpls-tp-identifiers [5]. The Type is 0x1. The length is 0x8.

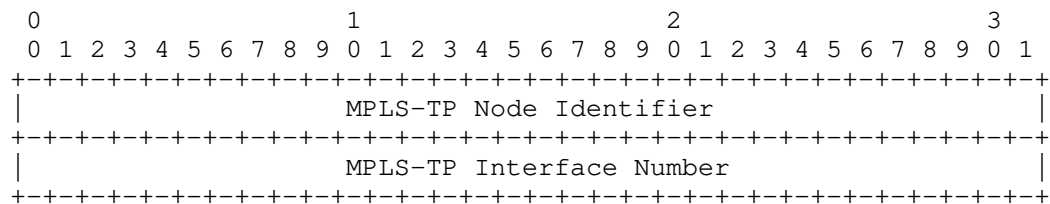


Figure 5: Interface Identifier TLV Format

#### 4.1.2. Global Identifier

This TLV carries the Interface Identifier as defined in draft-ietf-mpls-tp-identifiers [5]. The Type is 0x2. The length is 0x4.

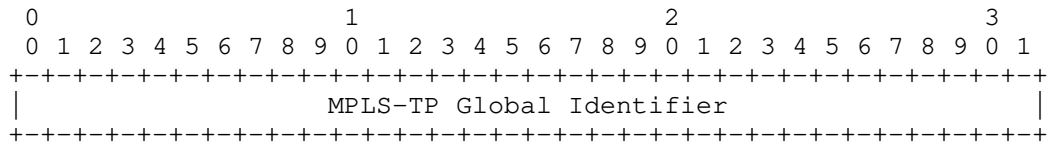


Figure 6: Global Identifier TLV Format

#### 4.1.3. International Carrier Code

This TLV carries the International Carrier Code. The Type is 0x3. The length is 0x8. The value is an Octet string padded with nulls.

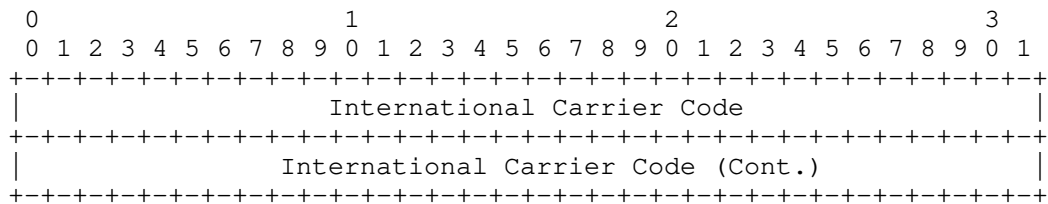


Figure 7: International Carrier Code

### 5. Sending and Receiving Fault Management Messages

#### 5.1. Sending a Fault Management Message

Service disruptive conditions are indicated by sending FM messages. The message type is set to the value corresponding to the condition. The refresh timer is set to the maximum time between successive FM messages. This value MUST not be changed on successive FM messages. If the optional clearing procedures are not used, then the default value is 1. Otherwise the default value is 20.

A Global-ID TLV or an ICC TLV MAY be included. The IF\_ID TLV SHOULD be included. If the R-Flag clearing procedures are to be used, the IF\_ID TLV MUST be included.

The message is then sent. The message MUST be refreshed two more times at an interval of one second. Further refreshes are sent according to the value of the refresh timer. Refreshing continues until the condition is cleared.

## 5.2. Clearing a FM Indication

Ceasing to send FM messages will clear the indication after 3.5 times the Refresh Timer. To clear an indication more quickly, the following procedure is used. The R-Flag of the FM message is set to one. Other fields of the FM message SHOULD NOT be modified. The message is sent immediately and then refreshed two more times at an interval of one second.

## 5.3. Receiving a FM Indication

When a FM message is received, a MEP examines it to ensure that that it is well formed. If the message type is unknown, the message is ignored. If the R-Flag is zero, the condition corresponding to the message type is entered. A timer is set to 3.5 times the refresh timer. If the message is not refreshed within this period, the condition is cleared. A message is considered a refresh if the message type and IF\_ID match an existing condition and the R-Flag is set to zero.

If the R-Flag is set to one, the MEP checks to see if a condition matching the message type and IF\_ID exists. If it does, that condition is cleared. Otherwise the message is ignored.

## 6. Minimum Implementation Requirements

At a minimum an implementation MUST support the following:

1. Sending AIS and LKR messages at a rate of 1 per second. In particular other values of the Refresh Timer and setting the R bit to value other than zero need not be supported.
2. Support of the sending the LDI flag.
3. Receiving AIS and LKR messages with any allowed Refresh Timer value.

The following items are optional to implement.

1. Support of receiving the LDI flag.
2. Support of receiving the R flag.
3. All TLVs.

## 7. Security Considerations

Spurious fault OAM messages form a vector for a denial of service attack. However, since these messages are carried in a control channel, one would have to gain access to a node providing the service in order to effect such an attack. Since transport networks are usually operated as a walled garden, such threats are less likely.

## 8. IANA Considerations

### 8.1. Pseudowire Associated Channel Type

Fault OAM requires a unique Associated Channel Type which are assigned by IANA from the Pseudowire Associated Channel Types Registry.

Registry:

Value	Description	TLV Follows	Reference
0xHHHH	Fault OAM	No	(This Document)

### 8.2. MPLS Fault OAM Message Type Registry

This sections details the MPLS Fault OAM TLV Registry, a new name spaces to be managed by IANA. The Type space is divided into assignment ranges; the following terms are used in describing the procedures by which IANA allocates values: "Standards Action" (as defined in RFC 5226 [6]) and "Private Use".

MPLS Fault OAM Message Types take values in the range 0-255. Assignments in the range 0-251 are via Standards Action; values in the range 251-255 are for Private Use, and MUST NOT be allocated.

Message Types defined in this document are:

Msg Type	Description
0x0	Reserved
0x1	Alarm Indication Signal (AIS)

0x2

Lock Report (LKR)

### 8.3. MPLS Fault OAM TLV Registry

This section details the MPLS Fault OAM TLV Registry, a new name space to be managed by IANA. The Type space is divided into assignment ranges; the following terms are used in describing the procedures by which IANA allocates values: "Standards Action" (as defined in RFC 5226 [6]), "Specification Required" and "Private Use".

MPLS Fault OAM TLVs which take values in the range 0-255. Assignments in the range 0-191 are via Standards Action; assignments in the range 192-248 are made via "Specification Required"; values in the range 248-255 are for Private Use, and MUST NOT be allocated.

TLVs defined in this document are:

Value	TLV Name
-----	-----
0	Reserved
1	Interface Identifier TLV
2	Global Identifier
3	International Carrier Code

## 9. References

### 9.1. Normative References

- [1] Niven-Jenkins, B., Brungard, D., Betts, M., Sprecher, N., and S. Ueno, "Requirements of an MPLS Transport Profile", RFC 5654, September 2009.
- [2] Vigoureux, M., Ward, D., and M. Betts, "Requirements for Operations, Administration, and Maintenance (OAM) in MPLS Transport Networks", RFC 5860, May 2010.
- [3] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [4] Bocci, M., Vigoureux, M., and S. Bryant, "MPLS Generic Associated Channel", RFC 5586, June 2009.
- [5] Bocci, M., Swallow, G., and E. Gray, "MPLS-TP Identifiers", draft-ietf-mpls-tp-identifiers-03 (work in progress), October 2010.
- [6] Narten, T. and H. Alvestrand, "Guidelines for Writing an IANA

Considerations Section in RFCs", BCP 26, RFC 5226, May 2008.

## 9.2. Informative References

### Authors' Addresses

George Swallow (editor)  
Cisco Systems, Inc.  
300 Beaver Brook Road  
Boxborough, Massachusetts 01719  
United States

Email: swallow@cisco.com

Annamaria Fulignoli (editor)  
Ericsson

Email: annamaria.fulignoli@ericsson.com

Martin Vigoureux (editor)  
Alcatel-Lucent  
Route de Villejust  
Nozay, 91620  
France

Email: martin.vigoureux@alcatel-lucent.com

Sami Boutros  
Cisco Systems, Inc.  
3750 Cisco Way  
San Jose, California 95134  
USA

Email: sboutros@cisco.com

David Ward  
Juniper Networks, Inc.

Email: dward@juniper.net

Stewart Bryant  
Cisco Systems, Inc.  
250, Longwater  
Green Park, Reading RG2 6GB  
UK

Email: [stbryant@cisco.com](mailto:stbryant@cisco.com)

Siva Sivabalan  
Cisco Systems, Inc.  
2000 Innovation Drive  
Kanata, Ontario K2K 3E8  
Canada

Email: [msiva@cisco.com](mailto:msiva@cisco.com)



Network Working Group  
Internet-Draft  
Intended status: Standards Track  
Expires: April 25, 2011

L. Jin  
ZTE  
F. Jounay  
France Telecom  
I. Wijnands  
Cisco Systems  
N. Leymann  
Deutsche Telekom AG  
October 22, 2010

Multicast LDP extension for hub & spoke multipoint LSP  
draft-jin-jounay-mpls-mlbp-hsmp-01.txt

Abstract

This draft introduces a hub & spoke multipoint LSP (short for HSMP LSP), which allows traffic both from root to leaf through P2MP LSP and also leaf to root along the co-routed reverse path. That means traffic entering the HSMP LSP from application/customer at the root node travels downstream, exactly as if it was traveling downstream along a P2MP LSP to each leaf node, and traffic entering the HSMP LSP at any leaf node travels upstream along the tree to the root. A packet traveling upstream should be thought of as being unicast to the root, except that it follows the path of the tree rather than ordinary unicast path.

Status of this Memo

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

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

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

This Internet-Draft will expire on April 25, 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.

## Table of Contents

1. Introduction . . . . .	3
2. Applications . . . . .	3
3. Terminology . . . . .	4
4. Setting up HSMP LSP with LDP . . . . .	4
4.1. Support for HSMP LSP setup with LDP . . . . .	5
4.2. HSMP FEC Elements . . . . .	5
4.3. Using the HSMP FEC Elements . . . . .	5
4.3.1. HSMP LSP Label Map . . . . .	6
4.3.2. HSMP LSP Label Withdraw . . . . .	8
4.3.3. HSMP LSP upstream LSR change . . . . .	8
5. HSMP LSP on a LAN . . . . .	8
6. Redundancy considerations . . . . .	9
7. Security Considerations . . . . .	9
8. IANA Considerations . . . . .	9
9. Acknowledgement . . . . .	9
10. References . . . . .	10
10.1. Normative references . . . . .	10
10.2. Informative References . . . . .	10
Authors' Addresses . . . . .	11

## 1. Introduction

The point-to-multipoint LSP defined in [I-D. draft-ietf-mpls-ldp-p2mp] allows traffic to transmit from root to several leaf nodes, and multipoint-to-multipoint LSP allows traffic from every node to transmit to every other node. This draft introduces a hub & spoke multipoint LSP (short for HSMP LSP), which allows traffic both from root to leaf through P2MP LSP and also leaf to root along the co-routed reverse path. That means traffic entering the HSMP LSP at the root node travels downstream, exactly as if it was traveling downstream along a P2MP LSP, and traffic entering the HSMP LSP at any other node travels upstream along the tree to the root. A packet traveling upstream should be thought of as being unicast to the root, except that it follows the path of the tree rather than ordinary unicast path.

## 2. Applications

There are applications that require such kind of LDP based HSMP LSP. According to time synchronization described in [IEEE1588v2], the sync packet and delay request should follow the same path, so as to provide same transmission delay for the two kinds of packets. By using point-to-multipoint technology to transmit these packets will greatly improve the bandwidth usage for above applications. Unfortunately current point-to-multipoint LSP only provides unidirectional path from source to leaf, which cannot fulfill the above new requirement. The main motivation of this draft is to solve the new problem. LDP based HSMP LSP described in this draft provides co-routed reverse path from leaf to root based on current unidirectional point-to-multipoint LSP.

There are two main specific scenarios for timing synchronization based on [IEEE1588v2]: 1. HSMP for phase/time delivery with TCKs. 2. HSMP for phase/time delivery with BCKs. The benefit of using mLDLP based HSMP LSP here is to provision dynamically the topology.

Time synchronization is required for accurate quantification of one-way delay as described in [I-D. draft-ietf-mpls-tp-loss-delay]. HSMP LSP can be used to do time synchronization based on [IEEE1588v2] for P2MP LSP or P2MP PW.

The mLDLP based HSMP LSP can also be applied in a typical IPTV scenario. There is usually only one location with senders but there are many receiver locations. If IGMP is used for signaling between senders and receivers, the messages from the receivers are travelling only from the leaves to the root (and from root towards leaves) but not from leaf to leaf. In addition traffic from the root is only

replicated towards the leaves. Then leaf node receiving IGMP message (for SSM case) will join HSMP LSP, and send IGMP message upstream to root along HSMP LSP.

Point to multipoint PW described in [I-D. draft-ietf-pwe3-p2mp-pw] requires to setup reverse path from leaf node (referred as egress PE) to root node (referred as ingress PE), if HSMP LSP is used to multiplex P2MP PW, the reverse path can also be multiplexed to HSMP upstream path to avoid setup independent reverse path. In that case, the operational cost will be reduced for maintaining only one HSMP LSP, instead of P2MP LSP and n (number of leaf nodes) P2P reverse LSPs.

### 3. Terminology

mLDP: Multicast LDP.

P2MP LSP: An LSP that has one Ingress LSR and one or more Egress LSRs.

MP2MP LSP: An LSP that connects a set of nodes, such that traffic sent by any node in the LSP is delivered to all others.

HSMP LSP: hub & spoke multipoint LSP. An LSP allows traffic both from root to leaf through P2MP LSP and also leaf to root along the co-routed reverse path.

### 4. Setting up HSMP LSP with LDP

HSMP LSP is similar with MP2MP LSP described in [I-D. draft-ietf-mpls-ldp-p2mp], with the difference that the leaf LSRs can only send traffic to root node along the same path of traffic from root node to leaf node.

HSMP LSP consists of a downstream path and upstream path. The downstream path is same as MP2MP LSP, while the upstream path is only from leaf to root node, without communication between leaf and leaf nodes. The transmission of packets from the root node of a HSMP LSP to the receivers is identical to that of a P2MP LSP. Traffic from a leaf node follows the upstream path toward the root node, along the identical path of downstream path.

For setting up the upstream path of a HSMP LSP, ordered mode MUST be used which is same as MP2MP. Ordered mode can guarantee a leaf to start sending packets to root immediately after the upstream path is installed, without being dropped due to an incomplete LSP.

Due to much of same behavior between HSMP LSP and MP2MP LSP, the following sections only describe the difference between the two entities.

#### 4.1. Support for HSMP LSP setup with LDP

HSMP LSP also needs the LDP capabilities [RFC5561] to indicate the supporting for the setup of HSMP LSPs. An implementation supporting the HSMP LSP procedures specified in this document MUST implement the procedures for Capability Parameters in Initialization Messages. Advertisement of the HSMP LSP Capability indicates support of the procedures for HSMP LSP setup.

A new Capability Parameter TLV is defined, the HSMP LSP Capability. Following is the format of the HSMP LSP Capability Parameter.

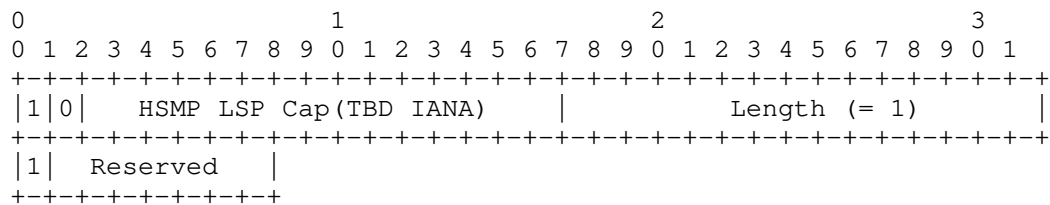


Figure 1

The HSMP LSP capability type is to be assigned by IANA.

#### 4.2. HSMP FEC Elements

Similar as MP2MP LSP, we define two new protocol entities, the HSMP downstream FEC and upstream FEC Element. Both elements will be used as FEC Elements in the FEC TLV. The structure, encoding and error handling for the HSMP downstream and upstream FEC Elements are the same as for the MP2MP FEC Element described in [I-D. draft-ietf-mpls-ldp-p2mp] Section 4.2. The difference is that two additional new FEC types are used: HSMP downstream type (TBD, IANA) and HSMP upstream type (TBD, IANA).

#### 4.3. Using the HSMP FEC Elements

In order to describe the message processing clearly, following defines the processing of the HSMP FEC Elements, which is inherited from [I-D. draft-ietf-mpls-ldp-p2mp] section 4.3.

1. HSMP downstream LSP <X, Y> (or simply downstream <X, Y>): a HSMP

LSP downstream path with root node address X and opaque value Y.

2. HSMP upstream LSP <X, Y> (or simply upstream <X, Y>): a HSMP LSP upstream path for root node address X and opaque value Y which will be used by any of downstream node to send traffic upstream to root node.

3. HSMP downstream FEC Element <X, Y>: a FEC Element with root node address X and opaque value Y used for a downstream HSMP LSP.

4. HSMP upstream FEC Element <X, Y>: a FEC Element with root node address X and opaque value Y used for an upstream HSMP LSP.

5. HSMP-D Label Map <X, Y, L>: A Label Map message with a single HSMP downstream FEC Element <X, Y> and label TLV with label L. Label L MUST be allocated from the per-platform label space of the LSR sending the Label Map Message.

6. HSMP-U Label Map <X, Y, Lu>: A Label Map message with a single HSMP upstream FEC Element <X, Y> and label TLV with label Lu. Label Lu MUST be allocated from the per-platform label space of the LSR sending the Label Map Message.

#### 4.3.1. HSMP LSP Label Map

This section specifies the procedures for originating HSMP Label Map messages and processing received HSMP label map messages for a particular HSMP LSP. The procedure of downstream HSMP LSP is same as that of downstream MP2MP LSP described in [I-D. draft-ietf-mpls-ldp-p2mp]. Under the operation of ordered mode, the upstream LSP will be setup by sending HSMP LSP mapping message with label which is allocated by upstream LSR to its downstream LSR one by one from root to leaf node, installing the upstream forwarding table by every node along the LSP. Detail procedure of upstream HSMP LSP is different with that of upstream MP2MP LSP, and is specified in below section.

All labels discussed here are downstream-assigned [RFC5332] except those which are assigned using the procedures described in section 5.

Determining the upstream LSR for a HSMP LSP <X, Y> follows the procedure for a MP2MP LSP described in [I-D. draft-ietf-mpls-ldp-p2mp] Section 4.3.1.1.

Determining one's downstream HSMP LSR procedure is much same as defined in [I-D. draft-ietf-mpls-ldp-p2mp] section 4.3.1.2. A LDP peer U which receives a HSMP-D Label Map from a LDP peer D will treat D as downstream HSMP LSR.

Determining the forwarding interface to an LSR has same procedure as defined in [I-D. draft-ietf-mpls-ldp-p2mp] section 2.4.1.2.

#### 4.3.1.1. HSMP LSP leaf node operation

The leaf node operation is same as the operation of MP2MP LSP defined in [I-D. draft-ietf-mpls-ldp-p2mp] section 4.3.1.4, only with different FEC element processing and specified below.

A leaf node Z will send a HSMP-D Label Map <X, Y, L> to U, instead of MP2MP-D Label Map <X, Y, L>. and expects a HSMP-U Label Map <X, Y, Lu> from node U and checks whether it already has forwarding state for upstream <X, Y>. The created forwarding state on leaf node Z is same as the leaf node of MP2MP LSP. Z will push label Lu onto the traffic that Z wants to forward over the HSMP LSP.

#### 4.3.1.2. HSMP LSP transit node operation

Suppose node Z receives a HSMP-D Label Map <X, Y, L> from LSR D, the procedure is same as processing MP2MP-D Label Mapping message defined in [I-D. draft-ietf-mpls-ldp-p2mp] section 4.3.1.5, and the processing protocol entity is HSMP-D label mapping message. The different procedure is specified below.

Node Z checks if upstream LSR U already assigned a label Lu to upstream <X, Y>. If not, transit node Z waits until it receives a HSMP-U Label Map <X, Y, Lu> from LSR U. Once the HSMP-U Label Map is received from LSR U, node Z checks whether it already has forwarding state upstream <X, Y> with incoming label Lu' and outgoing label Lu. If it does, Z sends a HSMP-U Label Map <X, Y, Lu'> to downstream node. If it does not, it allocates a label Lu' and creates a new label swap for Lu' with Label Lu over interface Iu. Interface Iu is determined via the procedures in Section 4.3.1. Node Z determines the downstream HSMP LSR as per Section 4.3.1, and sends a HSMP-U Label Map <X, Y, Lu'> to node D.

Since a packet from any downstream node is forwarded only to the upstream node, the same label (representing the upstream path) can be distributed to all downstream nodes. This differs from the procedures for MPMP LSPs [I-D. draft-ietf-mpls-ldp-p2mp], where a distinct label must be distributed to each downstream node. The forwarding state upstream <X, Y> on node Z will be like this {<Lu'>, <Iu Lu>}. Iu means the upstream interface over which Z receives HSMP-U Label Map <X, Y, Lu> from LSR U. Packets from any downstream interface over which Z send HSMP-U Label Map <X, Y, Lu'> with label Lu' will be forwarded to Iu with label Lu' swap to Lu.

#### 4.3.1.3. HSMP LSP root node operation

Suppose root node Z receives a HSMP-D Label Map <X, Y, L> from node D, the procedure is much same as processing MP2MP-D Label Mapping message defined in [I-D. draft-ietf-mpls-ldp-p2mp] section 4.3.1.6, and the processing protocol entity is HSMP-D label mapping message. The different procedure is specified below.

Node Z checks if it has forwarding state for upstream <X, Y>. If not, Z creates a forwarding state for incoming label Lu' that indicates that Z is the LSP egress. E.g., the forwarding state might specify that the label stack is popped and the packet passed to some specific application. Node Z determines the downstream HSMP LSR as per section 4.3.1, and sends a HSMP-U Label Map <X, Y, Lu'> to node D.

Since Z is the root of the tree, Z will not send a HSMP-D Label Map and will not receive a HSMP-U Label Map.

#### 4.3.2. HSMP LSP Label Withdraw

The HSMP Label Withdraw procedure is much same as MP2MP leaf operation defined in [I-D. draft-ietf-mpls-ldp-p2mp] section 4.3.2, and the processing protocol entities are HSMP FECs. The only difference is process of HSMP-U label release message, which is specified below.

When a transit node Z receives a HSMP-U label release message from downstream node D, Z should check if there are any incoming interface in forwarding state upstream <X, Y>. If all downstream nodes are released and there is no incoming interface, Z should delete the forwarding state upstream <X, Y> and send HSMP-U label release message to its upstream node.

#### 4.3.3. HSMP LSP upstream LSR change

The procedure for changing the upstream LSR is the same as defined in [I-D. draft-ietf-mpls-ldp-p2mp] section 4.3.3, except it is applied to HSMP FECs.

### 5. HSMP LSP on a LAN

The procedure to process P2MP LSP on a LAN has been described in [I-D. draft-ietf-mpls-ldp-p2mp]. When the LSR forwards a packet downstream on one of those LSPs, the packet's top label must be the "upstream LSR label", and the packet's second label is "LSP label".

When establishing the downstream path of a HSMP LSP, as defined in [I-D.ietf-mpls-ldp-upstream], a label request for a LSP label is send to the upstream LSR. The upstream LSR should send label mapping that contains the LSP label for the downstream HSMP FEC and the upstream LSR context label. At the same time, it must also send label mapping for upstream HSMP FEC to downstream node. Packets sent by the upstream router can be forwarded downstream using this forwarding state based on a two label lookup. Packets traveling upstream need to be forwarded in the direction of the root by using the label allocated by upstream LSR.

## 6. Redundancy considerations

In some scenario, it is necessary to provide two root nodes for redundancy purpose. One way to implement this is to use two independent HSMP LSPs acting as active/standby. At one time, only one HSMP LSP will be active, and the other will be standby. After detecting the failure of active HSMP LSP, the root and leaf nodes will switch the traffic to the new active HSMP LSP which is switched from former standby LSP. The detail of redundancy mechanism will be for future study.

## 7. Security Considerations

The same security considerations apply as for the MP2MP LSP described in [I-D. draft-ietf-mpls-ldp-p2mp].

## 8. IANA Considerations

This document requires allocation of two new LDP FEC Element types:

1. the HSMP-upstream FEC type - requested value 0x09
2. the HSMP-downstream FEC type - requested value 0x10

This document requires the assignment of new code points for the Capability Parameter TLVs, corresponding to the advertisement of the HSMP LSP capabilities. The values requested are:

1. HSMP LSP Capability Parameter - requested value 0x050B

## 9. Acknowledgement

The author would like to thank Eric Rosen, Fei Su for their valuable

comments.

## 10. References

### 10.1. Normative references

- [I-D. draft-ietf-mpls-ldp-p2mp]  
Minei, I., Kompella, K., and I. Wijnands, "Label Distribution Protocol Extensions for Point-to-Multipoint and Multipoint-to-Multipoint Label Switched Paths", draft-ietf-mpls-ldp-p2mp (work in progress), October 2009.
- [I-D.ietf-mpls-ldp-upstream]  
Aggarwal, R. and J. Le Roux, "MPLS Upstream Label Assignment for LDP", draft-ietf-mpls-ldp-upstream-08 (work in progress), July 2010.
- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC5036] Andersson, L., Minei, I., and B. Thomas, "LDP Specification", RFC 5036 , October 2007.
- [RFC5332] Rosen, E. and R. Aggarwal, "MPLS Multicast Encapsulations", RFC5332 , June 2008.
- [RFC5561] Thomas, B., Raza, K., and S. Aggarwal, "LDP Capabilities", RFC5561 , July 2009.

### 10.2. Informative References

- [I-D. draft-ietf-mpls-tp-loss-delay]  
Frost, D. and S. Bryant, "Signaling Root-Initiated Point-to-Multipoint Pseudowires using LDP", draft-ietf-mpls-tp-loss-delay-00 (work in progress), July 2010.
- [I-D. draft-ietf-pwe3-p2mp-pw]  
Martini, L., Jounay, F., Vecchio, G., Delord, S., Jin, L., and L. Ciavaglia, "Signaling Root-Initiated Point-to-Multipoint Pseudowires using LDP", draft-ietf-pwe3-p2mp-pw-00 (work in progress), July 2010.
- [IEEE1588v2]  
"IEEE standard for a precision clock synchronization protocol for networked measurement and control systems", IEEE1588v2 , March 2008.

Authors' Addresses

Lizhong Jin  
ZTE Corporation  
889, Bibo Road  
Shanghai, 201203, China

Email: lizhong.jin@zte.com.cn

Frederic Jounay  
France Telecom  
2, avenue Pierre-Marzin  
22307 Lannion Cedex, FRANCE

Email: frederic.jounay@orange-ftgroup.com

IJsbrand Wijnands  
Cisco Systems, Inc  
De kleetlaan 6a  
Diegem 1831, Belgium

Email: ice@cisco.com

Nicolai Leymann  
Deutsche Telekom AG  
Winterfeldtstrasse 21  
Berlin 10781

Email: N.Leymann@telekom.de



Intended status: Informational  
Expires: January 11, 2011

July 12, 2010

Temporal and hitless path segment monitoring  
draft-koike-mpls-tp-temporal-hitless-psm-01.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 January 4, 2011.

Copyright Notice

Copyright (c) 2009 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.

## Abstract

The MPLS transport profile (MPLS-TP) is now being standardized to enable carrier-grade packet transport and complement converged packet network deployments. One of the most attractive features in MPLS-TP is OAM functions, which enable network operators or service providers to provide various maintenance characteristics, such as prompt fault location, substantial survivability, performance monitoring, and preliminary or in-service measurements, inherent in circuit-based transport networks.

One of the important characteristics which are common in transport network operation is the fault location. A segment monitoring function of a transport path is effective in terms of extension of the maintenance work and indispensable particularly when the OAM function is effective only between end points. However, the current approach for the segment monitoring (SPME) has some fatal issues. This document elaborates on the problem statement on the path segment monitoring function. Moreover, this document also requests to add network objectives to solve or improve the issues and proposes to reconsider the new improved method of the segment monitoring.

This document is a product of a joint Internet Engineering Task Force (IETF) / International Telecommunications Union Telecommunications Standardization Sector (ITU-T) effort to include an MPLS Transport Profile within the IETF MPLS and PWE3 architectures to support the capabilities and functionalities of a packet transport network.

## Table of Contents

1. Introduction.....	4
2. Conventions used in this document.....	4
2.1. Terminology.....	5
2.2. Definitions.....	5
3. Network objectives for monitoring.....	5
4. Problem statement.....	5
5. OAM functions for segment monitoring.....	8
6. Conclusion.....	9
7. Security Considerations.....	9
8. IANA Considerations.....	9
9. References.....	9
9.1. Normative References.....	9
9.2. Informative References.....	10
10. Acknowledgments.....	10

Authors' Addresses.....	10
-------------------------	----

## 1. Introduction

A packet transport network will enable carriers or service providers to use network resources efficiently and reduce network cost. For carrier-grade network operation, appropriate maintenance characteristics, such as prompt fault location, substantial survivability, performance monitoring, and preliminary or in-service measurement, are essential to ensure quality and reliability of a network. They are essential in transport networks and have evolved along with TDM, ATM, SDH, and OTN.

Unlike in SDH or OTN networks, where OAM is an inherent part of every frame and frames are also transmitted in idle mode, it is not per se possible to constantly monitor the status of individual connections in packet networks. Packet-based OAM functions are flexible and selectively configurable to operator needs.

According to the MPLS-TP OAM requirements [1], mechanisms MUST be available for alerting a service provider of a fault or defect affecting the service(s) it provides. In addition, to ensure that faults or degradations can be localized, operators need a method to analyze or investigate the problem. From the fault localization perspective, end-to-end monitoring or management is not enough as a maintenance tool. In end-to-end OAM monitoring, when one problem occurs in an MPLS-TP network, the operator can detect the fault, but cannot solve the issue efficiently. As the operator cannot localize the point, he/she has no choice but to thoroughly search by trial and error by replacing packages or changing configuration by disrupting the connection service provided to customers. This is very inefficient and far from a carrier-grade network.

As a result, a specific segment monitoring function for detailed analysis, by focusing on and selecting a specific portion of a transport path, is indispensable to promptly and efficiently localize the fault point. This is an important characteristic in transport networks. However, a few fatal missing features, which are normally met in transport network operation, were found regarding the segment monitoring function of a transport path in the on-going MPLS-TP OAM standard.

This document elaborates on the problem statement on the path segment monitoring function. Moreover, this document also requests to add network objectives to solve or improve the issues and proposes to reconsider the new improved method of the segment monitoring.

## 2. Conventions used in 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 RFC-2119 [1].

## 2.1. Terminology

LSP Label Switched Path

OTN Optical Transport Network

PST Path Segment Tunnel

TCM Tandem connection monitoring

SDH Synchronous Digital Hierarchy

SPME Sub-path Maintenance Element

## 2.2. Definitions

None

## 3. Network objectives for monitoring

There are two missing and indispensable network objectives in the current on-going MPLS-TP standard.

(1) The monitoring and maintenance of current transport paths has to be conducted in service without traffic disruption.

(2) The monitored or managed transport path condition has to be exactly the same irrespective of any configurations necessary for maintenance.

It was agreed in ITU-T SG15 that Network objective (1) is mandatory and that regarding Network objective (2) the monitoring shall be hitless and not change the forwarding behavior.

## 4. Problem statement

To monitor, protect, or manage a portion of a transport path, such as LSP in MPLS-TP networks, the Sub-Path Maintenance Element (SPME) is defined in [2]. The SPME is defined between the edges of the portion of the LSP that needs to be monitored, protected, or managed. This SPME is created by stacking the shim header (MPLS header) [3] and is defined as the segment where the header is stacked. OAM messages can be initiated at the edge of the SPME and sent to the peer edge of the

SPME or to a MIP along the SPME by setting the TTL value of the label stack entry (LSE) and interface identifier value at the corresponding hierarchical LSP level in a per-node model.

This method can cause the following problems due to label stacking, which are fatal in terms of cost and operation.

(P-1) Increasing the bandwidth by the staking shim header(s)

(P-2) Increasing the address management of all MEPs and MIPs newly configured for SPME in the old MEG

Problem (P-1) increases the operation cost and wastes bandwidth. As the number of segments for monitoring increases, the number of label stacking increases. Moreover this is a permanent monitoring element, which should be set in advance, not a temporal setting. This prevents the operator from securing the shared bandwidth for efficient monitoring. However, the best solution is not to increase the bandwidth for monitoring.

Problem (P-2) is related to an identifier issue, of which the discussion is still continuing and not so fatal at the moment, but can be quite inefficient if the policy of allocating the identifier is independent in each layer. Moreover, from the perspective of operation, increasing the managed addresses and the managed layer is not desirable in terms of simplified operation featured by current transport networks. Reducing the managed identifier and managed layer should be the fundamental direction in designing the architecture.

The most familiar example of the SPME in the transport network standard is tandem connection monitoring (TCM), which is used for a carrier's carrier solution shown in Fig. 17 of the framework document[2]. However, in this case, the SPMEs have to be pre-configured. If this solution is applied to specific segment monitoring within one operator domain, all the necessary specific segments have to be pre-configured. This setting increases the managed objects as well as the necessary bandwidth, shown as Problem (P-1) and (P-2).

To avoid these issues, the temporal setting of the SPME(s) only when necessary seems reasonable and efficient for monitoring in MPLS-TP transport network operation. Unfortunately, the method of temporal settings of SPMEs also cause the following problems due to label stacking, which are fatal in terms of intrinsic monitoring and service disruption.

(P'-1) Changing the condition of the original transport path by changing the length of the MPLS frame (delay measurement and loss measurement can be sensitive)

(P'-2) Disrupting client traffic over a transport path, if the SPME is temporally configured.

Problem (P'-1) is a fatal problem in terms of intrinsic monitoring. The monitoring function checks the status without changing any conditions of the targeted monitored segment or the transport path. If the conditions of the transport path change, the measured value or observed data will also change. This can make the purpose of the monitoring meaningless because it seems very possible that the result of the monitoring does not reflect the data when the original fault or degradation occurred.

In addition, changing the settings of the original shim header should not be allowed because those changes correspond to creating a portion of the original transport path, which is completely different from the original circumstances.

Problem (P'-2) was not fully discussed, although the make-before-break procedure in the survivability document [4] seemingly supports the hitless configuration for monitoring according to the framework document [2]. The reality is the hitless configuration of SPME is impossible without affecting the circumstances of the targeted transport path, because the make-before-break procedure is premised on the change of the label value. This means changing one of the settings in MPLS shim header and should not be allowed as explained in (P'-1) above. Moreover, this is not effective under the static model without a control plane because the make-before-break is a restoration application based on the control plane.

To summarize, the problem statement is that the current sub-path maintenance based on a hierarchical LSP (SPME) is problematic in terms of increasing bandwidth by label stacking and managing objects by layer stacking and address management. A temporal configuration of SPME is one of the possible approaches for minimizing the impact of these issues; however, the current method is not applicable because the temporal configuration for monitoring can change the condition of the original monitored transport path and disrupt the in-service customer traffic. From the perspective of monitoring in transport network operation, the solution for avoiding those issues or minimizing their impact should be reconsidered.

## 5. OAM functions for segment monitoring

OAM functions in which segment monitoring is required are basically limited to on-demand monitoring which are defined in OAM framework document [5], because those segment monitoring functions are basically used to locate the fault/degraded point or diagnose the status for detailed analyses, especially when a problem occurred.

Packet loss and packet delay measurements are OAM functions in which hitless and temporal segment monitoring are strongly required because these functions are supported only between end points of a transport path. If a fault or defect occurs, there is no way to locate the fault or defect point without using the segment monitoring function. If an operator cannot locate or narrow the cause of the fault, it is quite difficult to take prompt action to solve the problem. Therefore, temporal and hitless monitoring for packet loss and packet delay measurements are indispensable for transport network operation.

Regarding other on-demand monitoring functions, segment monitoring is desirable, but not as urgent as the packet loss and packet delay measurements.

Regarding out-of-service on-demand monitoring functions, such as diagnostic tests, there is no need of hitless settings. However, specific segment monitoring is necessary for enabling the flexibility in diagnostic test patterns.

### Note:

The reason only on-demand OAM functions are discussed at this point is because the characteristic of "on-demand" is generally temporal for maintenance operation, and those operations are not reasonable and should not be based on pre-configuration. Pre-design and pre-configuration of PST/TCM (label stacking) for all the possible patterns for the on-demand (temporal) usage are not reasonable, because these tasks will additionally increase the operator's burden, although pre-configuration of PST for pro-active usage may be accepted considering the agreement thus far. Therefore, the solution for temporal and hitless segment monitoring does not need to be limited to label stacking mechanisms, such as PST/TCM(label stacking), which can cause the issues (P-1) and (P-2) described in Section 4.

Note that the issue of SPME is now under discussion in OAM framework draft [5]. The results of the discussion will have to be considered eventually. The solution for temporal and hitless

segment monitoring has to cover both per-node model and per-interface model in the draft [5].

## 6. Conclusion

We request that the above two network objectives in Section 3 are included in the MPLS-TP OAM framework. In addition, We also request that the solution for temporal and hitless path segment monitoring should never cause the problems mentioned in Section 4, i.e., P-1, P-2, P'-1 and P'-2, or at least minimize their impact to meet those two network objectives. As indicated liaison from ITU-T SG15, the provided solution is needed to become normative for the preparation of G.tpoam before it is consented.

## 7. Security Considerations

This document does not by itself raise any particular security considerations.

## 8. IANA Considerations

There are no IANA actions required by this draft.

## 9. References

### 9.1. Normative References

- [1] Vigoureux, M., Betts, M., Ward, D., "Requirements for OAM in MPLS Transport Networks", RFC5860, May 2010
- [2] Bocci, M., et al., "A Framework for MPLS in Transport Networks", RFC5921, July 2010
- [3] Rosen, E., et al., "MPLS Label Stack Encoding", RFC 3032, January 2001
- [4] Sprecher, N., Farrel, A. , "'Multiprotocol Label Switching Transport Profile Survivability Framework'", draft-ietf-mpls-tp-survive-fwk-05.txt(work in progress), April 2010
- [5] Busi, I., Niven-Jenkins, B. , "MPLS-TP OAM Framework", draft-ietf-mpls-tp-oam-framework-06.txt(work in progress), April 2010

## 9.2. Informative References

None

## 10. Acknowledgments

The author would like to thank all members (including MPLS-TP steering committee, the Joint Working Team, the MPLS-TP Ad Hoc Group in ITU-T) involved in the definition and specification of MPLS Transport Profile.

This document was prepared using 2-Word-v2.0.template.dot.

## Authors' Addresses

Yoshinori Koike (Editor)  
NTT  
Email: [koike.yoshinori@lab.ntt.co.jp](mailto:koike.yoshinori@lab.ntt.co.jp)



Internet Engineering Task Force  
Internet-Draft  
Intended status: Informational  
Expires: March 7, 2011

N. Leymann, Ed.  
Deutsche Telekom AG  
B. Decraene  
France Telecom  
C. Filsfils  
Cisco Systems  
D. Steinberg  
Steinberg Consulting  
September 3, 2010

Seamless MPLS Architecture  
draft-leymann-mpls-seamless-mpls-02

Abstract

This document describes an architecture which can be used to extend MPLS networks to integrate access and aggregation networks into a single MPLS domain ("Seamless MPLS"). The Seamless MPLS approach is based on existing and well known protocols. It provides a highly flexible and a scalable architecture and the possibility to integrate 100.000 of nodes. The separation of the service and transport plane is one of the key elements; Seamless MPLS provides end to end service independent transport. Therefore it removes the need for service specific configurations in network transport nodes (without end to end transport MPLS, some additional services nodes/configurations would be required to glue each transport domain). This draft defines a routing architecture using existing standardized protocols. It does not invent any new protocols or defines extensions to existing protocols.

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 March 7, 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.

## Table of Contents

1. Introduction . . . . .	4
1.1. Requirements Language . . . . .	4
1.2. Terminology . . . . .	4
2. Motivation . . . . .	5
2.1. Why Seamless MPLS . . . . .	6
2.2. Use Case #1 . . . . .	7
2.2.1. Description . . . . .	7
2.2.2. Typical Numbers . . . . .	10
2.3. Use Case #2 . . . . .	10
2.3.1. Description . . . . .	10
2.3.2. Typical Numbers . . . . .	12
3. Requirements . . . . .	12
3.1. Overall . . . . .	13
3.1.1. Access . . . . .	13
3.1.2. Aggregation . . . . .	13
3.1.3. Core . . . . .	14
3.2. Multicast . . . . .	14
3.3. Availability . . . . .	14
3.4. Scalability . . . . .	15
3.5. Stability . . . . .	15
4. Architecture . . . . .	15
4.1. Overall . . . . .	15
4.2. Multi-Domain MPLS networks . . . . .	15
4.3. Hierarchy . . . . .	16
4.4. Intra-Domain Routing . . . . .	16
4.5. Inter-Domain Routing . . . . .	16
4.6. Access . . . . .	17
5. Deployment Scenarios . . . . .	17
5.1. Deployment Scenario #1 . . . . .	17
5.1.1. Overview . . . . .	17

5.1.2.	General Network Topology . . . . .	17
5.1.3.	Hierarchy . . . . .	18
5.1.4.	Intra-Area Routing . . . . .	19
5.1.4.1.	Core . . . . .	19
5.1.4.2.	Aggregation . . . . .	19
5.1.5.	Access . . . . .	19
5.1.5.1.	LDP Downstream-on-Demand (DoD) . . . . .	20
5.1.6.	Inter-Area Routing . . . . .	21
5.1.7.	Labeled iBGP next-hop handling . . . . .	22
5.1.8.	Network Availability and Simplicity . . . . .	23
5.1.8.1.	IGP Convergence . . . . .	23
5.1.8.2.	Per-Prefix LFA FRR . . . . .	24
5.1.8.3.	Hierarchical Dataplane and BGP Prefix Independent Convergence . . . . .	24
5.1.8.4.	BGP Anycast . . . . .	25
5.1.8.5.	Applicability . . . . .	29
5.1.8.6.	Conclusion . . . . .	32
5.1.9.	Multicast . . . . .	33
5.1.10.	Next-Hop Redundancy . . . . .	33
5.2.	Scalability Analysis . . . . .	34
5.2.1.	Control and Data Plane State for Deployment Scenario #1 . . . . .	34
5.2.1.1.	Introduction . . . . .	34
5.2.1.2.	Core Domain . . . . .	35
5.2.1.3.	Aggregation Domain . . . . .	36
5.2.1.4.	Summary . . . . .	37
5.2.1.5.	Numerical application for use case #1 . . . . .	38
5.2.1.6.	Numerical application for use case #2 . . . . .	38
6.	Acknowledgements . . . . .	39
7.	IANA Considerations . . . . .	39
8.	Security Considerations . . . . .	40
9.	References . . . . .	40
9.1.	Normative References . . . . .	40
9.2.	Informative References . . . . .	40
Appendix A.	ABR Fast Reroute . . . . .	42
Authors' Addresses	. . . . .	45

## 1. Introduction

MPLS as a mature and well known technology is widely deployed in today's core and aggregation/metro area networks. Many metro area networks are already based on MPLS delivering Ethernet services to residential and business customers. Until now those deployments are usually done in different domains; e.g. core and metro area networks are handled as separate MPLS domains.

Seamless MPLS extents the core domain and integrates aggregation and access domains into a single MPLS domain ("Seamless MPLS"). This enables a very flexible deployment of an end to end service delivery. In order to obtain a highly scalable architecture Seamless MPLS takes into account that typical access devices (DSLAMs, MSAN) are lacking some advanced MPLS features, and may have more scalability limitations. Hence access devices are kept as simple as possible.

Seamless MPLS is not a new protocol suite but describes an architecture by deploying existing protocols like BGP, LDP and ISIS. Multiple options are possible and this document aims at defining a single architecture for the main function in order to ease implementation prioritization and deployments in multi vendor networks. Yet the architecture should be flexible enough to allow some level of personalization, depending on use cases, existing deployed base and requirements. Currently, this document focus on end to end unicast LSP.

### 1.1. Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [RFC2119].

### 1.2. Terminology

This document uses the following terminology

- o Access Node (AN): An access node is a node which processes customers frames or packets at Layer 2 or above. This includes but is not limited to DSLAMs or OLTs (in case of (G)PON deployments). Access nodes have only limited MPLS functionalities in order to reduce complexity in the access network.
- o Aggregation Node (AGN): An aggregation node (AGN) is a node which aggregates several access nodes (ANs).
- o Area Border Router (ABR): Router between aggregation and core domain.

- o **Deployment Scenario:** Describes which an implementation of Seamless MPLS in order to fullfil the requirements derived from one or more use cases.
- o **Seamless MPLS Domain:** A set of MPLS equipments which can set MPLS LSPs between them.
- o **Transport Node (TN):** Transport nodes are used to connect access nodes to service nodes, and services nodes to services nodes. Transport nodes ideally have no customer or service state and are therefore decoupled from service creation.
- o **Seamless MPLS (S-MPLS):** Used as a generic term to describe an architecture which integrates access, aggregation and core network in a single MPLS domain.
- o **Service Node (SN):** A service node is used to create services for customers and is connected to one or more transport nodes. Typical examples include Broadband Network Gateways (BNGs), video servers
- o **Transport Pseudo Wire (T-PW):** A transport pseudowire provides service independent transport mechanisms based on Pseudo-Wires within the Seamless MPLS architecture.
- o **Use Case:** Describes a typical network including service creation points in order to describe the requirments, typical numbers etc. which need to be taken into account when applying the Seamless MPLS architecture.

## 2. Motivation

MPLS is deployed in core and aggregation network for several years and provides a mature and stable basis for large networks. In addition MPLS is already used in access networks, e.g. such as mobile or DSL backhaul. Today MPLS as technology is being used on two different layers:

- o the Transport Layer and
- o the Service Layer (e.g. for MPLS VPNs)

In both cases the protocols and the encapsulation are identical but the use of MPLS is different especially concerning the signalling, the control plane, the provisioning, the scalability and the frequency of updates. On the service layer only service specific information is exchanged; every service can potentially deploy it's

own architecture and individual protocols. The services are running on top of the transport layer. Nevertheless those deployments are usually isolated, focussed on a single use case and not integrated into an end-to-end manner.

The motivation of Seamless MPLS is to provide an architecture which supports a wide variety of different services on a single MPLS platform fully integrating access, aggregation and core network. The architecture can be used for residential services, mobile backhaul, business services and supports fast reroute, redundancy and load balancing. Seamless MPLS provides the deployment of service creation points which can be virtually everywhere in the network. This enables network and service providers with a flexible service and service creation. Service creation can be done based on the existing requirements without the needs for dedicated service creation areas on fixed locations.

## 2.1. Why Seamless MPLS

Multiple SP plan to deploy networks with 10k to 100k MPLS nodes. This is typically at least one order of magnitude higher than typical deployments and may require a new architecture. Multiple options are possible and it makes sense for the industry (both vendors and SP) to restrict the options in order to ease the first deployments (e.g. restrict the number of options to implement and/or scales for vendors, reduce interoperability and debugging issues for SP).

Many aggregation networks are already deploying MPLS but are limited to the use of MPLS per aggregation area. Those MPLS based aggregation domains are connected to a core network running MPLS as well. Nevertheless most of the services are not limited to an aggregation domain but running between several aggregation domains crossing the core network. In the past it was necessary to provide connectivity between the different domains and the core on a per service level and not based on MPLS (e.g. by deploying native IP-Routing or Ethernet based technologies between aggregation and core). In most cases service specific configurations on the border nodes between core and aggregation were required. New services led to additional configurations and changes in the provisioning tools (see Figure 1).

With Seamless MPLS there are no technology boundaries and no topology boundaries for the services. Network (or region) boundaries are for scaling and manageability, and do not affect the service layer, since the Transport Pseudowire that carries packets from the AN to the SN doesn't care whether it takes two hops or twenty, nor how many region boundaries it needs to cross. The network architecture is about network scaling, network resilience and network manageability; the

service architecture is about optimal delivery: service scaling, service resilience (via replicated SNs) and service manageability. The two are decoupled: each can be managed separately and changed independently.

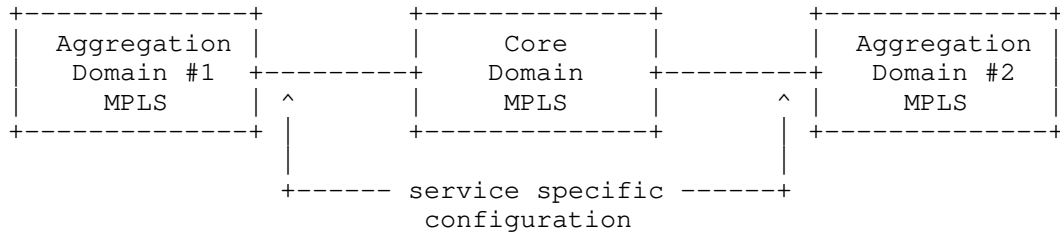


Figure 1: Service Specific Configurations

One of the main motivations of Seamless MPLS is to get rid of services specific configurations between the different MPLS islands. Seamless MPLS connects all MPLS domains on the MPLS transport layer providing a single transport layer for all services - independent of the service itself. The Seamless MPLS architecture therefore decouples the service and transport layer and integrates access, aggregation and core into a single platform. One of the big advantages is that problems on the transport layer only need to be solved once (and the solutions are available to all services). With Seamless MPLS it is not necessary to use service specific configurations on intermediate nodes; all services can be deployed in an end to end manner.

## 2.2. Use Case #1

### 2.2.1. Description

In most cases at least residential and business services need to be supported by a network. This section describes a Seamless MPLS use case which supports such a scenario. The use case includes point to point services for business customers as well as typical service creation for residential customers.

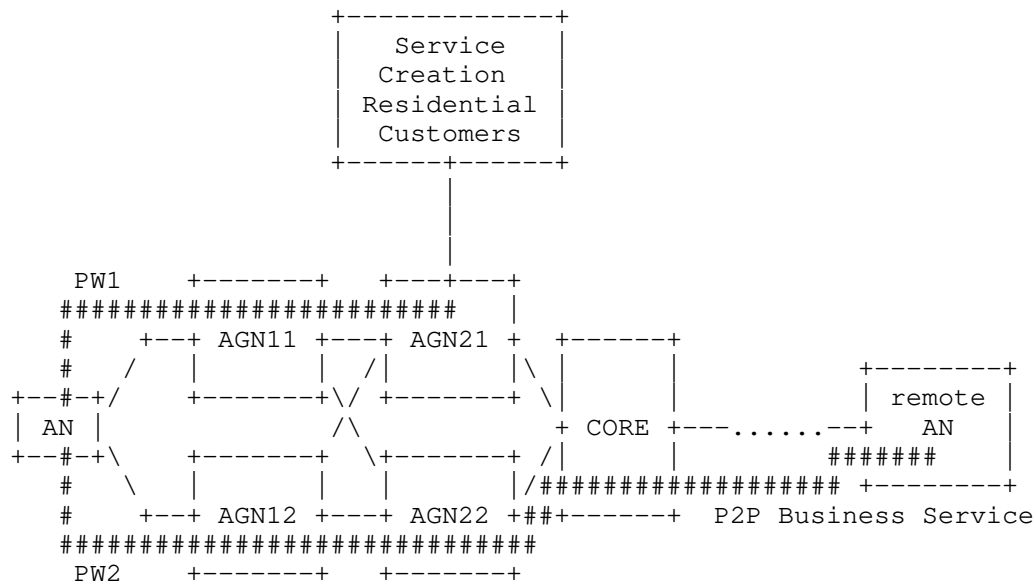


Figure 2: Use Case #1: Service Creation

Figure 2 shows the different service creation points and the corresponding pseudowires between the access nodes and the service creation points. The use case does not show all PWs (e.g. not the PWs needed to support redundancy) in order to keep the figure simple. Node and link failures are handled by rerouting the PWs (based on standard mechanisms).

**Residential Services:** The service creation for all residential customers connected to the Access Nodes in an aggregation domain is located on an Service Node connected to the AGN2x. The PW (PW1) originated at the AN and terminates at the AGN2. A second PW is deployed in the case where redundancy is needed on the AN (the figure shows redundancy but this might not be the case for all ANs in this Use Case). Additional PWs can be deployed as well in case more than a single service creation is needed for the residential service (e.g. one service creation point for Internet access and a second service creation point for IPTV services).

**Business Services:** For business services the use cases shows point to point connections between two access nodes. PW2 originates at the AN and terminates on the remote AN crossing two aggregation areas and the core network. If the access node needs connections to several remote ANs the corresponding number of PWs will be originated at the AN. Nevertheless taking the number of ports available and the number of business customers on a typical access

node the number of PWs will be relatively small.

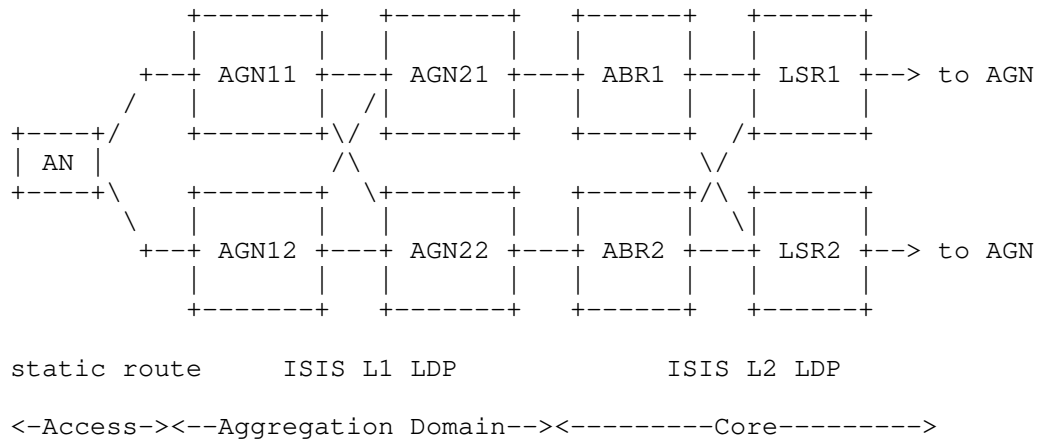


Figure 3: Use Case #1: Redundancy

Figure 3 shows the redundancy at the access and aggregation network deploying a two stage aggregation network (AGN1x/AGN2x). Nevertheless redundancy is not a MUST in this use case. It is also possible to use non redundant connection between the ANs and AGN1 stage and/or between the AGN1 and AGN2 stages. The AGN2x stage is used to aggregate traffic from several AGN1x pairs. In this use case an aggregation domain is not limited to the use of a single pair of AGN2x; the deployment of several AGN2 pairs within the domain is also supported. As design goal for the scalability of the routing and forwarding within the Seamless MPLS architecture the following numbers are used:

- o Number of Aggregation Domains: 100
- o Number of Backbone Nodes: 1.000
- o Number of Aggregation Nodes: 10.000
- o Number of Access Nodes: 100.000

The access nodes (AN) are dual homed to two different aggregation nodes (AGN11 and AGN12) using static routing entries on the AN. The ANs are always source or sink nodes for MPLS traffic but not transit nodes. This allows a light MPLS implementation in order to reduce the complexity in the AN. The aggregation network consists of two stages with redundant connections between the stages (AGN11 is connected to AGN21 and AGN22 as well as AGN12 to AGN21 and AGN22). The gateway between the aggregation and core network is realized

using the Area Border Routers (ABR). From the perspective of the MPLS transport layer all systems are clearly identified using the loopback address of the system. An ingress node must be able to establish a service to an arbitrary egress system by using the corresponding MPLS transport label

### 2.2.2. Typical Numbers

Table 1 shows typical numbers which are expected for Use Case #1 (access node).

Parameter	Typical Value
IGP Control Plane	2
IP FIB	2
LDP Control Plane	200
LDP FIB	200
BGP Control Plane	0
BGP FIB	0

Table 1: Use Case #1: Typical Numbers for Access Node

### 2.3. Use Case #2

#### 2.3.1. Description

In most cases, residential, wholesales and business services need to be supported by the network.

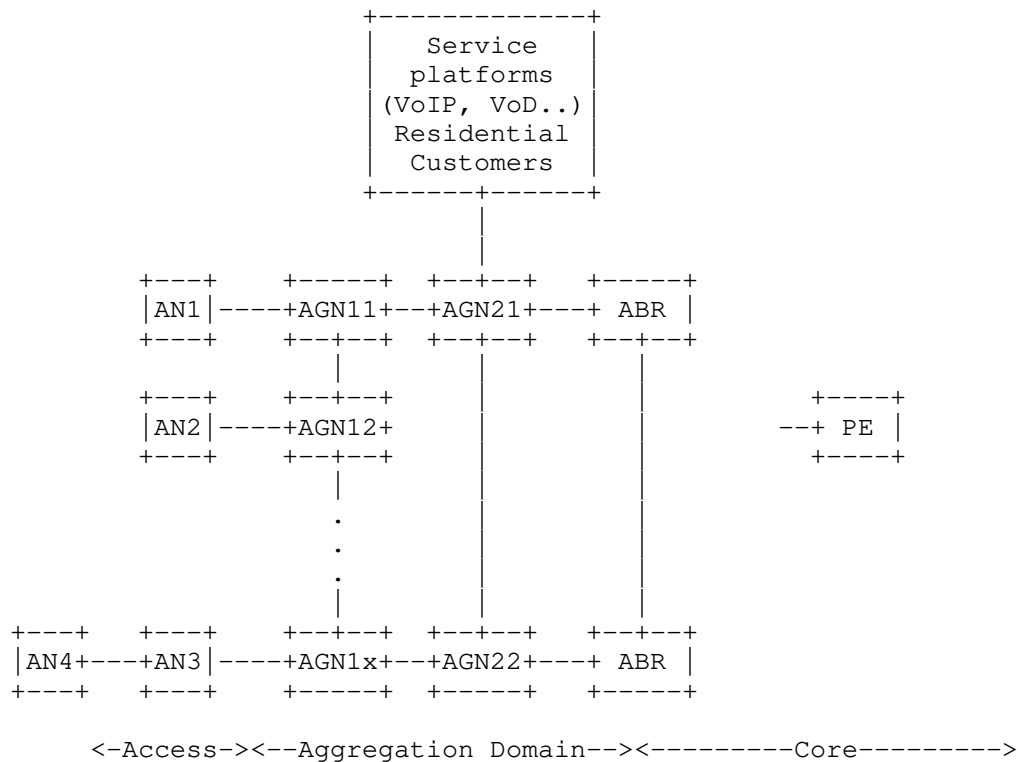


Figure 4: Use Case #2

The above topology (see Figure 4) is subject to evolutions, depending on AN types and capacities (in terms of number of customers and/or aggregated bandwidth). For examples, AGN1x connection toward AGN2y currently forms a ring but may latter evolve in a square or triangle topology; AGN2y nodes may not be present...

Most access nodes (AN) are single attached on one aggregation node using static routing entries on the AN and AGN. Some AN, are dual attached on two different AGN using static routes. Some AN are used as transit by some lower level AN. Static routes are expected to be used between those AN.

IPv4, IPv6 and MPLS interconnection between the aggregation and core network is realized using the Area Border Routers (ABR). Any ingress node must be able to establish IPv4, IPv6 and MPLS connections to any egress node in the seamless MPLS domain.

Regarding MPLS connectivity requirements, a full mesh of MPLS LSPs is required between the ANs of an aggregation area, at least for 6PE

purposes. Some additional LSPs are needed between ANs and some PE in the aggregation area or in the core area for access to services, wholesale and enterprises services. In short, a meshing of LSP is required between the AGN of the whole seamless MPLS domain. Finally, LSP between any node to any node should be possible.

From a scalability standpoint, the following numbers are the targets:

- o Number of Aggregation Domains: 30
- o Number of Backbone Nodes: 150
- o Number of Aggregation Nodes: 1.500
- o Number of Access Nodes: 40.000

### 2.3.2. Typical Numbers

Table 2 shows typical numbers which are expected for Use Case #2 for the purpose of establishing the transport LSPs. They do not take into account the services built in addition. (e.g. 6PE will require additional IPv6 routes).

Parameter	Typical Value
IGP Control Plane	2
IP FIB	2
LDP Control Plane	1000
LDP FIB	1000

Table 2: Use Case #2: Typical Numbers for Access Node

## 3. Requirements

The following section describes the overall requirements which need to be fulfilled by the Seamless MPLS architecture. Beside the general requirements of the architecture itself there are also certain requirements which are related to the different network nodes.

- o End to End Transport LSP: MPLS based services (pseudowire based, L3-VPN or IP) SHALL be provided by the Seamless MPLS based infrastructure between any nodes.

- o Scalability: The network SHALL be scalable to the minimum of 100.000 nodes.
- o Fast convergence (sub second resilience) SHALL be supported. Fast reroute (LFA) SHOULD be supported.
- o Flexibility: The Seamless MPLS architecture SHALL be applied to a wide variety of existing MPLS deployments. It SHALL use a flexible approach deploying building blocks with the possibility to use certain features only if those features are needed (e.g. dual homing ANs or fast reroute mechanisms).
- o Service independence: Service and transport layer SHALL be decoupled. The architecture SHALL remove the need for service specific configurations on intermediate nodes.
- o Native Multicast support: P2MP MPLS LSPs SHOULD be supported by the Seamless MPLS architecture.
- o Interoperable end to end OAM mechanisms SHALL be implemented

### 3.1. Overall

#### 3.1.1. Access

In respect of MPLS functionality the access network should be kept as simple as possible. Compared to the aggregation and/or core network within Seamless MPLS a typical access node is less powerful. The control plane and the forwarding should be as simple as possible. To reduce the complexity and the costs of an access node not the full MPLS functionality need to be supported (control and data plane). The use of an IGP should be avoided. Static routing should be sufficient. Required functionality to reach the required scalability should be moved out of the access node. The number of access nodes can be very high. The support of load balancing for layer 2 services should be implemented.

#### 3.1.2. Aggregation

The aggregation network aggregates traffic from access nodes. The aggregation Node must have functionalities that enlarge the scalability of the simple access nodes that are connected. The IGP must be link state based. Each aggregation area must be a separated area. All routes that are interarea should use an EGP to keep the IGP small. The aggregation node must have the full scalability concerning control plane and forwarding. The support of load balancing for layer 2 services must be implemented.

### 3.1.3. Core

The core connects the aggregation areas. The core network elements must have the full scalability concerning control plane and forwarding. The IGP must be link state based. The core area must not include routes from aggregation areas. All routes that are interarea should use an EGP to keep the IGP small. Each area of the link state based IGP should have less than 2000 routes. The support of load balancing for layer 2 services must be implemented.

### 3.2. Multicast

Compared with unicast connectivity Multicast is more dynamic. User generated messages - like joining or leaving multicast groups - are interacting directly with network components in the access and aggregation network (in order to build the corresponding forwarding states). This leads to the need for a highly dynamic handling of messages on access and aggregation nodes. Nevertheless the core network SHOULD be stable and state changes triggered by user generated messages SHOULD be minimized. This rises the need for an hierarchy for the P2MP support in Seamless MPLS hiding the dynamic behaviour of the access and aggregation nodes

- o mLDP
- o P2MP RSVP-TE

### 3.3. Availability

All network elements should be high available (99.999% availability). Outage times should be as low as possible. A repair time of 50 milliseconds or less should be guaranteed at all nodes and lines in the network that are redundant. Fast convergence features SHOULD be used in all control plane protocols. Local Repair functions SHOULD be used wherever possible. Full redundancy is required at all equipment that is shared in a network element.

- o Power Supply
- o Switch Fabric
- o Routing Processor

A change from an active component to a standby component SHOULD happen without effecting customers traffic. The Influence of customer traffic MUST be as low as possible.

### 3.4. Scalability

The network must be highly scalable. As a minimum requirement the following scalability figures should be met:

- o Number of aggregation domains: 100
- o Number of backbone nodes: 1.000
- o Number of aggregation nodes: 10.000
- o Number of access nodes: 100.000

### 3.5. Stability

- o The platform should be stable under certain circumstances (e.g. missconfiguration within one area should not cause instability in other areas).
- o Differentiate between "All Loopbacks and Link addresses should be ping able from every where." Vs. "Link addresses are not necessary ping able from everywhere".

## 4. Architecture

### 4.1. Overall

One of the key questions that emerge when designing an architecture for a seamless MPLS network is how to handle the sheer size of the necessary routing and MPLS label information control plane and forwarding plane state resulting from the stated scalability goals especially with respect to the total number of access nodes. This needs to be done without overwhelming the technical scaling limits of any of the involved nodes in the network (access, aggregation and core) and without introducing too much complexity in the design of the network while at the same time still maintaining good convergence properties to allow for quick MPLS transport and service restoration in case of network failures.

### 4.2. Multi-Domain MPLS networks

The key design paradigm that leads to a sound and scalable solution is the divide and conquer approach, whereby the large problem is decomposed into many smaller problems for which the solution can be found using well-known standard architectures.

In the specific case of seamless MPLS the overall MPLS network SHOULD

be decomposed into multiple MPLS domains, each well within the scaling limits of well-known architectures and network node implementations. From an organizational and operational point of view it MAY make sense to define the boundaries of such domains along the pre-existing boundaries of aggregation networks and the core network.

Examples of how networks can be decomposed include using IGP areas as well as using multiple BGP autonomous systems.

#### 4.3. Hierarchy

These MPLS domains SHOULD then be then be connected into an MPLS multi-domain network in a hierarchical fashion that enables the seamless exchange of loopback addresses and MPLS label bindings for transport LSPs across the entire MPLS internetwork while at the same time preventing the flooding of unnecessary routing and label binding information into domains or parts of the network that do not need them. Such a hierarchical routing and forwarding concept allows a scalability in different dimensions and allows to hide the complexity and size of the aggregation and access networks.

#### 4.4. Intra-Domain Routing

The intra-domain routing within each of the MPLS domains (i.e. aggregation domains and core) SHOULD utilize standard IGP protocols like OSPF or ISIS. By definition, each of these domains is small enough so that there are no relevant scaling limits within each IGP domain, given well-known state-of-the-art IGP design principles and recent router technology.

The intra-domain MPLS LSP setup and label distribution SHOULD utilize standard protocols like LDP or RSVP.

#### 4.5. Inter-Domain Routing

The inter-domain routing is responsible for establishing connectivity between and across all MPLS domains. The inter-domain routing SHOULD establish a routing and forwarding hierarchy in order to achieve the scaling goals of seamless MPLS. Note that the IP aggregation usually performed between region (IGP areas/AS) in IP routing does not work for MPLS as MPLS is not capable of aggregating FEC (because MPLS forwarding use an exact match lookup, while IP uses longest match).

Therefore it is RECOMMENDED to utilize protocols that support indirect next-hops (like BGP with MPLS labels "labeled BGP/SAFI4" [RFC3107]).

#### 4.6. Access

Compared to the aggregation and core parts of the Seamless MPLS network the access part is special in two respects:

- o The number of nodes in the access is at least one order of magnitude higher than in any other part of the network.
- o Because of the large quantity of access nodes, the cost of these nodes is extremely relevant for the overall costs of the entire network, i.e. access nodes are very cost sensitive.

This makes it desirable to design the architecture such that the AN functionality can be kept as simple as possible. This should always be kept in mind when evaluating different seamless MPLS architectures. The goal is to limit both the number of different protocols needed on the AN as well as the scale to which each protocol must perform to the absolute minimum.

#### 5. Deployment Scenarios

This section describes the deployment scenarios based on the use cases and the generic architecture above.

##### 5.1. Deployment Scenario #1

Section describing the Seamless MPLS implementation of a large european ISP.

###### 5.1.1. Overview

This deployment scenario describes one way to implement a seamless MPLS architecture. Specific to this implementation is the choice of intra- and inter-domain routing and label distribution protocols, as well as the details of the interworking of these protocols to achieve the overall scalable hierarchical architecture.

###### 5.1.2. General Network Topology

There are multiple aggregation domains (in the order of up to 100) connected to the core in a star topology, i.e. aggregation domains are never connected among themselves, but only to the core. The core has its own domain.

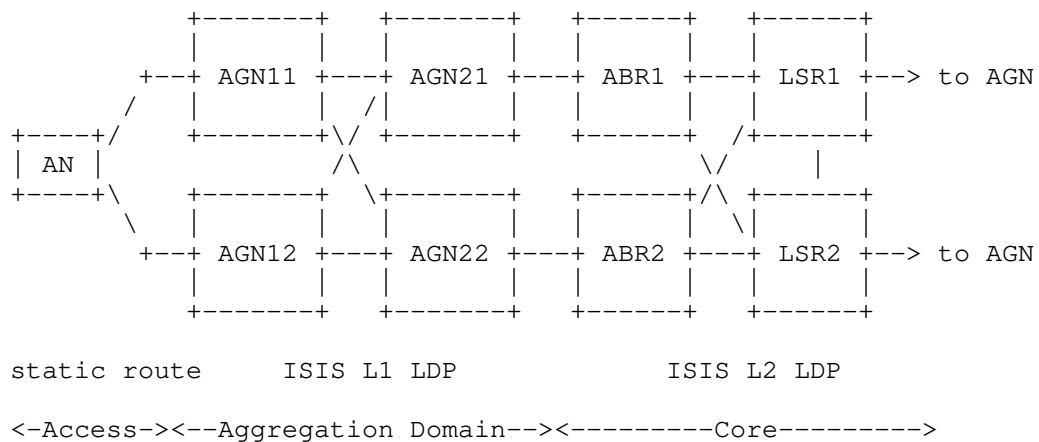


Figure 5: Deployment Scenario #1

As shown in Figure 5, the access nodes (AN) are connected to the aggregation network via aggregation nodes called AGN1x, either to a single AGN1x or redundantly to two AGN1x. Each AGN1x has redundant uplinks to a pair of second-level aggregation nodes called AGN2x.

Each aggregation domain is connected to the core via exactly two border routers (ABR) on the core side. There can be multiple AGN2 pairs per aggregation domain, but only one ABR pair for each aggregation domain. Each of the AGN2 in an AGN2 pair connects to one of the ABRs in the ABR pair responsible for that aggregation domain.

The ABRs on the core side have redundant connections to a pair of LSR routers.

The LSR pair is also connected via a direct link.

The core LSR are connected to other core LSR in a partly meshed topology so that there are disjunct, redundant paths from each LSR to each other LSR.

### 5.1.3. Hierarchy

As explained before, hierarchy is the key to a scalable seamless MPLS architecture. The hierarchy in this implementation is achieved by forming different MPLS domains for aggregation domains and core, where within each of these domains a fairly common MPLS deployment using ISIS as intradomain link-state routing protocol and using LDP for MPLS label distribution is used.

These MPLS domains are mapped to ISIS areas as follows: Aggregation

domains are mapped to ISIS L1 areas. The core is configured as ISIS L2. The border routers connecting aggregation and core are ISIS L1L2 and are referred to as ABRs. From a technical and operational point of view these ABRs are part of the core, although they also belong to the respective aggregation domain purely from a routing protocol point of view.

For the interdomain-routing BGP with MPLS labels is deployed ("labeled BGP/SAFI4" [RFC3107]).

#### 5.1.4. Intra-Area Routing

##### 5.1.4.1. Core

The core uses ISIS L2 to distribute routing information for the loopback addresses of all core nodes. The border routers (ABR) that connect to the aggregation domains are also part of the respective aggregation ISIS L1 area and hence ISIS L1L2.

LDP is used to distribute MPLS label binding information for the loopback addresses of all core nodes.

##### 5.1.4.2. Aggregation

The aggregation domains uses ISIS L1 as intra-domain routing protocol. All AGN loopback addresses are carried in ISIS.

As in the core, the aggregation also uses LDP to distribute MPLS label bindings for the loopback addresses.

##### 5.1.5. Access

Access nodes do not have their own domain or IGP area. Instead, they directly connect to the AGN1 nodes in the aggregation domain. To keep access devices as simple as possible, ANs do not participate in ISIS.

Instead, each AN has two static default routes pointing to each of the AGN1 it is connected to. Appropriate techniques SHOULD be deployed to make sure that a given default route is invalidated when the link to an AGN1 or that node itself fails. Examples of such techniques include monitoring the physical link state for loss of light/loss of frame, or using Ethernet link OAM or BFD [I-D.ietf-bfd-v4v6-1hop].

The AGN1 MUST have a configured static route to the loopback address of each of the ANs it is connected to, because it cannot learn the AN loopback address in any other way. These static routes have to be

monitored and invalidated if necessary using the same techniques as described above for the static default routes on the AN.

The AGN1 redistributes these routes into ISIS for intra-domain reachability of all AN loopback addresses.

LDP is used for MPLS label distribution between AGN1 and AN. In order to keep the AN control plane as lightweight as possible, and to avoid the necessity for the AN to store 100.000 MPLS label bindings for each upstream AGN1 peer, LDP is deployed in downstream-on-demand (DoD) mode, described below.

To allow the label bindings received via LDP DoD to be installed into the LFIB on the AN without having the specific host route to the destination loopback address, but only a default route, use of the LDP Extension for Inter-Area Label Switched Paths [RFC5283] is made.

#### 5.1.5.1. LDP Downstream-on-Demand (DoD)

LDP downstream-on-demand mode is specified in [RFC5036]. Although it was originally intended to be used with ATM switch hardware, there is nothing from a protocol perspective preventing its use in a regular MPLS frame-based environment. In this mode the upstream LSR will explicitly ask the downstream LSR for a label binding for a particular FEC when needed.

The assumption is that a given AN will only have a limited number of services configured to an even more limited number of destinations, or egress LER. Instead of learning and storing all label bindings for all possible loopback addresses within the entire Seamless MPLS network, the AN will use LDP DoD to only request the label bindings for the FECs corresponding to the loopback addresses of those egress nodes to which it has services configured.

For LDP DoD the AGN1 MUST also ask the AN for label bindings for specific FECs. FECs are necessary for all pseudowire destinations at the AN. Most preferable this pseudowire destination is the LSR-ID of the AN. Depending on the AN implementation and architecture multiple pseudowire destination addresses and associated FECs could be needed. The conclusion of this results to the following requirement:

- o The AGN1 MUST ask the AN for label bindings for all potential pseudowire destination addresses on the AN. Because the AGN (at least in many cases) does not take part in the pseudowire signaling an independent way of receiving the AN FEC is necessary on the AGN. These potential pseudowire destinations MUST be known on the AGN1, by configuration or otherwise. These are typically the loopback addresses of the AN, to which a static route has been

configured anyway on the AGN1, as explained above. In addition to these static routes, the AGN1 SHOULD be configured statically to request MPLS label bindings for these loopback addresses via LDP DoD.

- o Optionally an automatism that asks for a FEC for the LSR-ID COULD be implemented. A configuration switch that disables this option must be implemented. The label is necessary. The way of initiating the DoD-signaling of the label could be done with both methods (configuration/automatism).
- o The AN knows by configuration to which destination a pseudowire is set up. The AN is always the endpoint of the pseudowire. Before signalling a pseudowire the AN MUST ask (via LDP DoD) the AGN for a FEC. Because of this an independent preconfiguration is not necessary on the AN.
- o The following are the triggers for ANs to request a label:
  - o
    - \* When a control session (targeted LDP) to a target has to be established
    - \* When a service label has been received by a control session (e.g. pseudo wire label)

#### 5.1.6. Inter-Area Routing

The inter-domain MPLS connectivity from the aggregation domains to and across the core domain is realized primarily using BGP with MPLS labels ("labeled BGP/SAFI4" [RFC3107]). A very limited amount of route leaking from ISIS L2 into L1 is also used.

All ABR and PE nodes in the core are part of the labeled iBGP mesh, which can be either full mesh or based on route reflectors. These nodes advertise their respective loopback addresses (which are also carried in ISIS L2) into labeled BGP.

Each ABR node has labeled iBGP sessions with all AGN1 nodes inside the aggregation domain that they connect to the core. Since there are two ABR nodes per aggregation domain, this leads to each AGN1 node having an iBGP sessions with each of the two ABR. Note that the use of iBGP implies that the entire seamless MPLS internetwork is just a single AS to which all core and aggregation nodes belong. The AGN1 nodes advertise their own loopback addresses into labeled BGP, in addition to these loopbacks also being in ISIS L1.

Additionally the AGN1 nodes also redistribute all the statically configured routes to the AN loopback addresses into labeled BGP. Note that as stated above, the AGN1 MUST ask the AN for label bindings for the AN loopback FECs via LDP DoD in order to have a valid labeled route with a non-null label.

This architecture results in carrying all loopbacks of all nodes except pure P nodes (AN, AGN, ABR and core PE) in labeled BGP, e.g. there will be in the order of 100.000 routes in labeled BGP when approaching the stated scalability goal. Note that this only affects the BGP RIB size and does not necessarily imply that any node needs to actually have active forwarding state (LFIB) in the same order of magnitude. In fact, as will be discussed in the scalability analysis, no single node needs to install all labeled BGP routes into the LFIB, but each node only needs a small percentage of the RIB as active forwarding state in the LFIB. And from a RIB point of view, BGP is known to scale to hundreds of thousands of routes.

#### 5.1.7. Labeled iBGP next-hop handling

The ABR nodes run labeled iBGP both to the core mesh as well as to the AGN1 nodes of their respective aggregation domains. Therefore they operate as iBGP route reflectors, reflecting labeled routes from the aggregation into the core and vice versa.

When reflecting routes from the core into the aggregation domain, the ABR SHOULD NOT change then BGP NEXT-HOP addresses (next-hop-unchanged). This is the usual behaviour for iBGP route reflection. In order to make these routes resolvable to the AGN1 nodes inside the aggregation domain, the ABR MUST leak all other ABR and core PE loopback addresses from ISIS L2 into ISIS L1 of the aggregation domain. Note that the number of leaked addresses is limited so that the overall scalability of the seamless MPLS architecture is not impacted. In the worst case all core loopback addresses COULD be leaked into ISIS L1, but even that would not be a scalability problem.

When reflecting routes from the aggregation into the core, the ABR MUST set then BGP NEXT-HOP to its own loopback addresses (next-hop-self). This is not the default behaviour for iBGP route reflection, but requires special configuration on the ABR. Note that this also implies that the ABR MUST allocate a new local MPLS label for each labeled iBGP FEC that it reflects from the aggregation into the core. This special next-hop handling is essential for the scalability of the overall seamless MPLS architecture since it creates the required hierarchy and enables the hiding of all aggregation and access addresses behind the ABRs from an IGP point of view. Leaking of aggregation ISIS L1 loopback addresses into ISIS L2 is not necessary

and MUST NOT be allowed.

he resulting hierarchical inter-domain MPLS routing structure is similar to the one described in [RFC4364] section 10c, only that we use one AS with route reflection instead of using multiple ASes.

#### 5.1.8. Network Availability and Simplicity

The seamless mpls architecture illustrated in deployment case study 1 guarantees a sub-second loss of connectivity upon any link or node failures. Furthermore, in the vast majority of cases, the loss of connectivity is limited to sub-50msec.

These network availability properties are provided without any degradation on scale and simplicity. This is a key achievement of the design.

In the remaining of this section, we first introduce the different network availability technologies and then review their applicability for each possible failure scenario.

##### 5.1.8.1. IGP Convergence

IGP convergence can be modelled as a linear process with an initial delay and a linear FIB update [ACM01].

The initial delay could conservatively be assumed to be 260msec: 50msec to detect failures with BFD (most failures would be detected faster with loss of light for example or with faster BFD timers), 50msec to throttle the LSP generation, 150msec to throttle the SPF computation (making sure than all the required LSP's are received even in case of SRLG failures) and 10msec for shortest-path-first tree computation.

Assuming 250usec per update (conservative), this allows for  $(1000-260)/0.250 = 2960$  prefixes update within a second following the outage. More precisely, this allows for 2960 important IGP prefixes updates. Important prefixes are automatically classified by the router implementation through simple heuristic (/32 is more important than non-/32).

The number of IGP important routes (loopbacks) in deployment case study 1 is much smaller than 2960, and hence sub-second IGP convergence is conservative.

IGP convergence is a simple technology for the operator provided that the router vendor optimizes the default IGP behavior (no need to tune any arcane knob).

#### 5.1.8.2. Per-Prefix LFA FRR

A per-prefix LFA for a destination D is a precomputed backup IGP nexthop for that destination. This backup IGP nexthop can be link protecting or node protecting [RFC5286].

The analysis of the applicability of Per-Prefix LFA in the deployment model 1 of Seamless MPLS architecture is straightforward thanks to [I-D.filsfils-rtgwg-lfa-applicability].

In deployment model 1, each aggregation network either follows the triangle or full-mesh topology. Further more, the backbone region implements a dual-plane. As a consequence, the failure of any link or node within an aggregation domain is protected by LFA FRR (sub-50msec) for all impacted IGP prefixes, whether intra-area or inter-area. No uloop may form as a result of these failures [I-D.filsfils-rtgwg-lfa-applicability].

Per-Prefix LFA FRR is generally assessed as a simple technology for the operator [I-D.filsfils-rtgwg-lfa-applicability]. It certainly is in the context of deployment case study 1 as the designer enforced triangle and full-mesh topologies in the aggregation network as well as a dual-plane core network.

#### 5.1.8.3. Hierarchical Dataplane and BGP Prefix Independent Convergence

In a hierarchical dataplane, the FIB used by the packet processing engine reflects the recursions between routes. For example, a BGP route B recursing on IGP route I whose best path is via interface O is encoded as a FIB entry B pointing to a FIB entry I pointing to a FIB entry O.

BGP PIC [BGPPIC] extends the hierarchical dataplane with the concept of a BGP Path-List. A BGP path-list may be abstracted as a set of primary multipath nhops and a backup nhop. When the primary set is empty, packets destined to the BGP destinations are rerouted via the backup nhop.

With BGP PIC and hierarchical dataplane, a FIB entry representing a BGP route points to a FIB entry representing a BGP Path-List. This entry may either point again to another BGP Path list entry (BGP over BGP recursion) or more likely points to a FIB entry representing an IGP route.

A BGP Path-list may be computed automatically by the router and does not require any operator involvement. Specifically, the automated computation adapts to any routing policy (this is key to understand the simplicity of BGP PIC and the ability to enable it as a default

router behavior). There is no constraint at all on the operator design. Any policy is supported (multipath, primary/backup between neighboring domains or via alternate domains).

The BGP backup nhop is computed in advance of any failure (ie. a second bestpath computation after excluding the primary nhops).

Hierarchical dataplane and BGP PIC provide two important routing availability properties.

First, upon IGP convergence, recursive BGP routes immediately benefit from the updated IGP paths thanks to the dataplane indirection. This is key as most of the traffic is destined to BGP routes, not to IGP routes.

Second, upon loss of the primary BGP nhop, the dataplane can immediately reroute the packets towards the pre-computed backup nhop. This redirection is said to be prefix independent as the only entries that need to be modified are the BGP path-lists. These entries are shared across all the BGP prefixes with the same primary and backup next-hops. This scale independence is key. In the context of deployment model 1, while there might be 100k BGP routes, we only expect on the order of 200 BGP path-lists. Assuming 10usec in-place modification per BGP path-list, we see that the router can enable the backup path for 100k BGP destinations in less than 2msec (less than  $200 * 10\text{usec}$ ).

The detection of the loss of the primary BGP nhop (and hence the need to enable the pre-computed backup BGP nhop) can be local (a local link failing between an edge device and a single-hop eBGP peer) or involves an IGP convergence (a remote border router goes down).

These BGP PIC properties benefit to any BGP routes: Internet, L3VPN, 3107, IPv4 or IPv6. Future evolution of VPLS will also benefit from such properties [I-D.raggarwa-mac-vpn], [I-D.sajassi-l2vpn-rvpls-bgp]

Hierarchical forwarding and BGP PIC are very simple technology to operate. Their ability to adapt to any topology, any routing policy and any BGP address family allows router vendors to enable this behavior by default.

#### 5.1.8.4. BGP Anycast

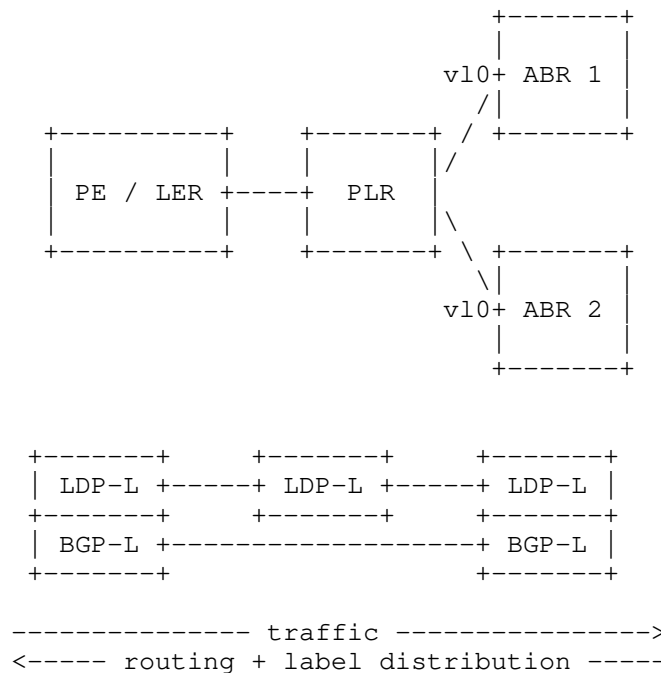


Figure 6: Routing and Traffic Flow

In order to deploy local repair also when an area border router fails an additional mechanism is used (see Figure 6) with the PLR being the point of local repair. ABR1 advertises in addition to it's loopback address in ISIS and LDP a virtual loopback address (vl0). ABR2 is backup router and advertises the same address as well with a worse metric. If ABR1 fails the PLR can immediately locally switch to the alternative path towards ABR2 (anycast behaviour). If this anycast address is used as next hop within the labeled BGP a local repair for the MPLS transport layer can be performed.

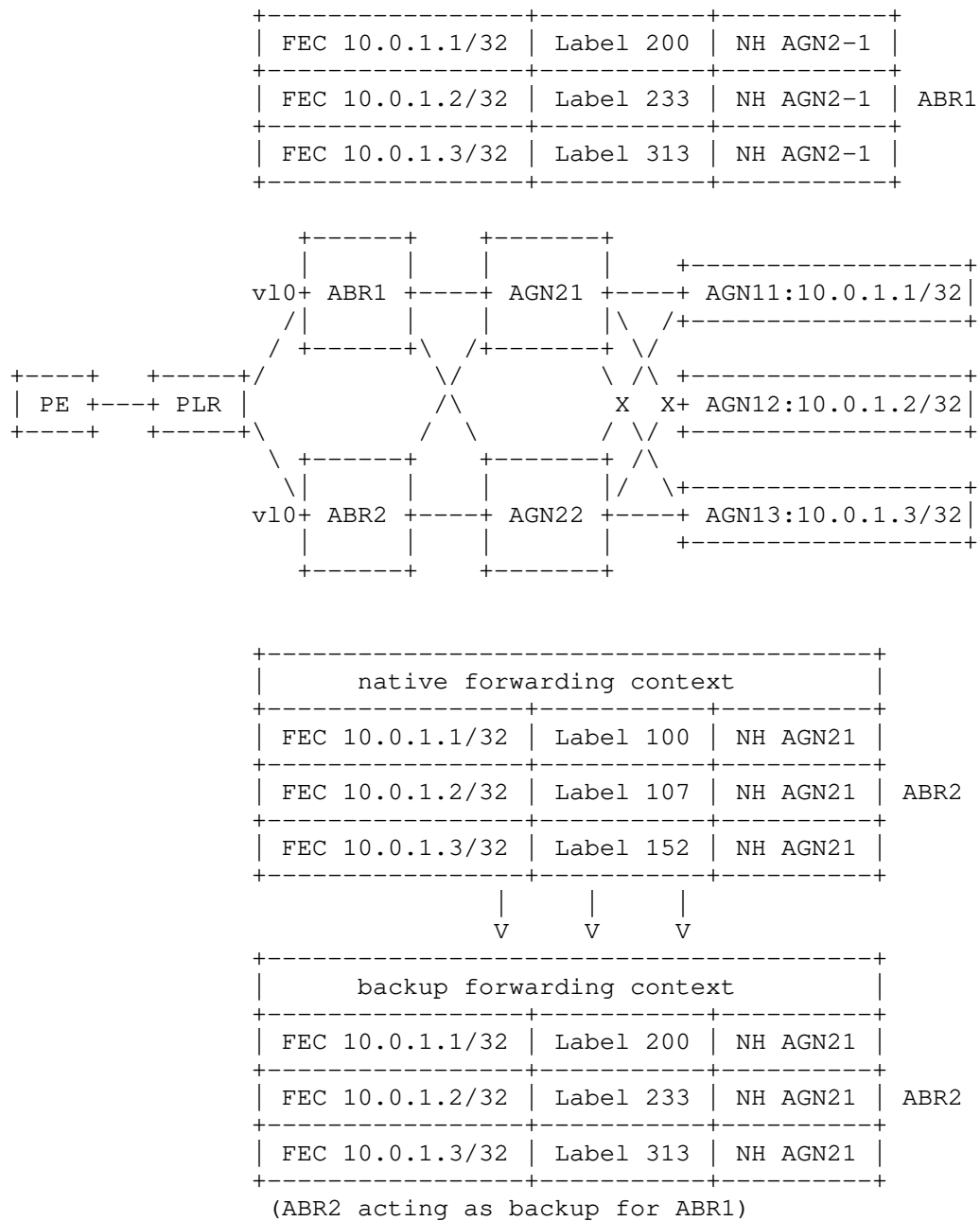


Figure 7: ABR Failure Scenarios

Figure 7 shows the behaviour in case of an ABR failure. ABR2 creates

a backup forwarding context for labeled BGP routes which are routed via ABR1. This local backup forwarding context is not part of the global label table. ABR2 advertises over LDP as anycast next-hop not implicit null but an explicit label which points internally to the backup context. ABR2 learns via BGP the labels used by ABR1 for the connected aggregation domain(s). In addition ABR2 carries also label forwarding states for the corresponding FECs. By combining those two informations the backup forwarding context is filled with valid label forwarding entries. MPLS frames carrying the correct label for the forwarding by ABR1 as second entry in the label stack are also correctly forwarded by ABR2 using this information.

The behavior described above allows for a local protection upon ABR node failure, hence a <50msec protection. In the same scenario, BGP PIC would provide a <600msec protection as the prefix-independent switch-over on the pre-computed backup nhop occurs upon IGP convergence (deletion of the IGP route to the remote ABR is <600msec).

ABR node protection is only required on the ABR's while BGP PIC technology is required on all the AGN1 routers.

Deploying and operating BGP anycast has significant complexity implications.

First, the scheme is applicable in deployment case study one because the two ABR's have the same exact connectivity. This assumption is not met by most deployment case studies. Indeed, in the context of OSPF deployments, it frequently occurs that ABR's are connected to different areas. For example, ABR1 is connected to area 1, 2, 3 and 4 while ABR2 is connected to ABR 3, 4, 5 and 6. The reachability between ABR's is not equivalent. Upon loss of ABR1, it would be incorrect to redirect all the ABR1-bound traffic to ABR2 as ABR2 cannot handle traffic to areas 1 and 2. Enabling BGP anycast in this case leads to blackhole.

Second, the anycast solution is very specific to the topology of deployment model 1. In this very specific (and rare) topology each neighbor of an ABR is also a neighbor of the mirrored ABR. It is only in this very specific topology that an LFA construction can be created with anycast addressing.

Third, any misconfiguration of the anycast loopback address may create blackhole (the traffic is attracted to a destination that does not know about a route to the final destination).

As a consequence, this scheme cannot be generalized. It cannot handle any PE node failure in any topology and any routing policy.

Furthermore, it does not apply to PE-CE/Peering link failures.

BGP PIC's applicability is much wider than ABR local protection. BGP PIC applies to any PE node failure in any topology and any routing policy. Furthermore, it does apply to PE-CE/Peering link failures.

#### 5.1.8.5. Applicability

We select two typical traffic flows and analyze the loss of connectivity (LoC) upon each possible failure.

Flow F1 starts from an AN1 in a left aggregation region and ends on an AN2 in a right aggregation region. Each AN is dual-homed to two AGN's.

Flow F2 starts from an L3VPN PE1 in the core and ends at an L3VPN PE2 in the core.

##### 5.1.8.5.1. AN1-AGN link failure or AGN node failure

F1 is impacted but LoC <50msec is possible assuming fast BFD detection and fast-switchover implementation on the AN. F2 is not impacted.

##### 5.1.8.5.2. Link or node failure within the left aggregation region

F1 is impacted but LoC <50msec thanks to LFA FRR. No uloop will occur during the IGP convergence following the LFA protection. Note: if LFA is not available (other topology than case study one) or if LFA is not enabled, then the LoC would be < second as the number of impacted important IGP route in a seamless architecture is much smaller than 2960.

F2 is not impacted.

##### 5.1.8.5.3. ABR node failure between left region and the core

F1 is impacted but LoC <50msec thanks to LFA FRR. No uloop will occur during the IGP convergence following the LFA protection.

Note: This case is also called "Local ABR failure" as the ABR which fails is the one connected to the aggregation region at the source of flow F1.

Note: remember that the left region receives the routes to all the remote ABR's and that the labelled BGP routes are reflected from the core to the left region with next-hop unchanged. This ensures that the loss of the (local) ABR between the left region and the core is

seen as an IGP route impact and hence can be addressed by LFA.

Note: if LFA is not available (other topology than case study one) or if LFA is not enabled, then the LoC would be < second as the number of impacted important IGP route in a seamless architecture is much smaller than 2960.

F2 is not impacted.

#### 5.1.8.5.4. Link or node failure within the core region

F1 and F2 are impacted but LoC <50msec thanks to LFA FRR.

This is specific to the particular core topology used in deployment case study 1. The core topology has been optimized [I-D.filsfils-rtgwg-lfa-applicability] for LFA applicability.

As explained in [I-D.filsfils-rtgwg-lfa-applicability], another alternative to provide <50msec in this case consists in using an MPLS-TE full-mesh and MPLS-TE FRR. This is required when the designer is not able or does not want to optimize the topology for LFA applicability and he wants to achieve <50msec protection.

Alternatively, simple IGP convergence would ensure a LoC < second as the number of impacted important IGP route in a seamless architecture is much smaller than 2960.

#### 5.1.8.5.5. PE2 failure

F1 is not impacted.

F2 is impacted and the LoC is sub-300msec thanks to IGP convergence and BGP PIC.

The detection of the primary nhop failure (PE2 down) is performed by a single-area IGP convergence.

In this specific case, the convergence should be much faster than <sec as very few prefixes are impacted upon an edge node failure. Reusing the introduction on IGP convergence presented in an earlier section and assuming 2 important impacted prefixes (two loopbacks per edge node), one would expect that PE2's failure is detected in 260msec + 2\*0.250msec.

In a hierarchical FIB organization, once the loss of a primary BGP nhop is detected, it only takes a few msec's to update all the impacted BGP Path-Lists and hence divert the impacted BGP/VPN traffic towards the precomputed backup nhops.

The LoC for BGP/BPN traffic upon PE2 failure is thus expected to be <300msec.

#### 5.1.8.5.6. PE2's PE-CE link failure

F1 is not impacted.

F2 is impacted and the LoC is sub-50msec thanks to local interface failure detection and BGP PIC.

#### 5.1.8.5.7. ABR node failure between right region and the core

F2 is not impacted.

F1 is impacted. We analyze the LoC for F1 for both BGP PIC and BGP anycast.

LoC is sub-600msec thanks to BGP PIC.

The detection of the primary nhop failure (ABR down) is performed by a multi-area IGP convergence.

First, the two (local) ABR's between the left and core regions must complete the core IGP convergence. The analysis is similar to the loss of PE2. We would thus expect that the core convergence completes in ~260msec.

Second, the IGP convergence in the left region will cause all AGN1 routers to detect the loss of the remote ABR. This second IGP convergence is very similar to the first one (2 important prefixes to remove) and hence should also complete in ~260msec.

Once an AGN1 has detected the loss of the remote ABR, thanks to hierarchical FIB organization, in-place modification of shared BGP path-list and pre-computation of BGP backup nhop, the AGN1 reroutes flow F1 via the alternate remote ABR in a few msec's.

As a consequence, the LoC for F1 upon remote ABR failure is thus expected to be <600msec.

Provided that all the strict topological assumptions have been met and the additional operational complexity is deemed justifiable, LoC is sub-50msec with BGP anycast .

#### 5.1.8.5.8. Link or node failure within the right aggregation region

F1 is impacted but LoC <50msec thanks to LFA FRR. No uloop will occur during the IGP convergence following the LFA protection.

Note: if LFA is not available (other topology then case study one) or if LFA is not enabled, then the LoC would be < second as the number of impacted important IGP route in a seamless architecture is much smaller than 2960.

F2 is not impacted.

#### 5.1.8.5.9. AGN (connected to AN2) node failure

F1 is impacted but LoC <50msec thanks to LFA FRR. No uloop will occur during the IGP convergence following the LFA protection.

Note: remember that AGN redistributes the static routes to ANs within ISIS. The loss of an AGN on the IGP path to AN2 is thus seen as an IGP route impact and hence LFA FRR is applicable.

Note: if LFA is not available (other topology then case study one) or if LFA is not enabled, then the LoC would be < second as the number of impacted important IGP route in a seamless architecture is much smaller than 2960.

F2 is not impacted.

#### 5.1.8.5.10. AGN-AN2 link failure

F2 is not impacted.

F1 is impacted.

LoC is sub-300msec with IGP convergence as only one prefix needs to be updated.

Sub-50msec could be guaranteed provided that the LFA implementation supports a redistributed static as a native IGP route.

#### 5.1.8.5.11. AN2 failure

F1 is impacted and the LoC lasts until the AN is recovered.

F2 is not impacted.

#### 5.1.8.6. Conclusion

The Seamless MPLS architecture illustrated in deployment case study 1 guarantees sub-50msec upon any link or node failures.

These properties are provided by LFA FRR and BGP PIC and are applicable to any traffic flow, AN to AN, AN to SPE or L3VPN PE to

L3VPN PE.

The availability properties can be characterized as scale-independent as the number of IGP important routes is independent of the number of AN's and the LFA FRR and BGP PIC technologies are prefix-independent.

There are two exceptions: the loss of a remote ABR and the loss of a remote PE (PE2).

In both cases, IGP convergence and BGP PIC provide a simple (automated) sub-600msec protection.

In the case of the remote ABR failure, provided that all the strict topological assumptions have been met and the additional operational complexity to avoid misconfiguration and blackhole is deemed justifiable, LoC is sub-50msec with BGP anycast.

Note that this is not applicable to the remote PE failure.

#### 5.1.9. Multicast

#### 5.1.10. Next-Hop Redundancy

An aggregation domain is connected to the core network using two redundant area border routers, and MPLS hierarchy is applied on these ABRs. MPLS hierarchy helps scale the FIB but introduces additional complexity for the rerouting in case of ABR failure. Indeed ABR failure requires a BGP converge to update the inner MPLS hierarchy, in addition to the IGP converge to update the outer MPLS hierarchy. This is also expected to take more time as BGP convergence is performed after the IGP convergence and because the number of prefixes to update in the FIB can be significant. This is a drawback but the architecture allow for two "local" solutions which restore the traffic before the BGP convergence takes place.

One called BGP PIC edge, would be required on all edge LSR involved in the inner (BGP) MPLS hierarchy. Namely all routers except the AN which are not involved in the inner MPLS hierarchy. It involves pre-computing and pre-installing in the FIB the BGP backup path. Such back up path are activated when the IGP advertise the failure of the primary path.

One called egress fast reroute, would be required on the egress LSR involved in the inner (BGP) MPLS hierarchy, namely TN and AGN connected to ABR. It involves:

using a anycast loopback address shared by both nominal and back up ABR, advertised by both ABR in the IGP and advertised as BGP Next Hop by the nominal ABR;

activating IP FRR LFA on the (penultimate) hops, acting as PLR for the anycast loopback;

using on the backup egress nodes (ABR2) an additional contextual MPLS FIB populated by the labels upstream allocated by the nominal egress node (ABR1).

Details can be found in [PEFRR] and [ABRFRR], and in the appendix of this draft. Both solutions have their pro and con, and the choice is left to each Service Provider or deployment based on the different requirements. The point is that the seamless MPLS architecture can handles fast restoration time, even for ABR failures.

## 5.2. Scalability Analysis

### 5.2.1. Control and Data Plane State for Deployment Scenario #1

#### 5.2.1.1. Introduction

Let's call:

- o #AN the number of Access Node (AN) in the seamless MPLS domain
- o #AGN the number of AGgregation Node (AGN) in the seamless MPLS domain
- o #Core the number of Core (Core) in the core network
- o #Area the number of aggregation routing domains.

Let's take the following assumptions:

- o Aggregation equipments are equally spread across aggregation routing domains
- o the number of IGP links is three times the number of IGP nodes
- o the number of IGP prefixes is five times the number of IGP nodes (links prefixes + 2 loopbacks)
- o Access Nodes need to set up 1000 (1k) LSPs. 10% (100) are FEC which are outside of their routing domain. Those 100 remote FEC are the same for all Access Nodes of a given AGN.

The following sections roughly evaluate the scalability, both in absolute numbers and relatively with the number of Access Node which is the biggest scalability factor.

#### 5.2.1.2. Core Domain

The IGP & LDP core domain are not affected by the number of access nodes:

IGP:

node : #Core  $\sim o(1)$

links :  $3 * \#Core \sim o(1)$

IP prefixes :  $5 * \#Core \sim o(1)$

LDP FEC:

#Core  $\sim o(1)$

Core TN FIBs grows linearly with the number of node in the core domain. In other word, they are not affected by AGN and AN nodes:

Core TN:

IP FIB :  $5 * \#Core \sim o(1)$

MPLS LFIB : #Core  $\sim o(1)$

BGP carries all AN routes which is significant. However, all AN routes are only needed in the control plane, possibly in a dedicated BGP Route Reflector (just like for BGP/MPLS VPNs) and not in the forwarding plane. The number of routes (100k) is smaller than the number of number of routes in the Internet (300k and rising) or in major VPN SP (>500k and rising) so the target can be handled with current implementations. In addition, AN routes are internal routes whose churn and instability is smaller and more under control than external routes.

BGP Route Reflector (RR)

NLRI : #AN  $\sim o(n)$

path :  $2 * \#AN \sim o(2n)$

ABR handles both the core and aggregations routes. They do not depend on the total number of AN nodes, but only on the number of AN

in their aggregation domain.

ABR:

$$\text{IP FIB} : 5 * \text{Core} + (5 * \text{AGN} + \text{AN}) / \text{Area} \sim o(\text{AN} / \text{Area})$$

$$\text{MPLS LFIB} : \text{Core} + (\text{AGN} + \text{AN}) / \text{Area} \sim o(\text{AN} / \text{Area})$$

#### 5.2.1.3. Aggregation Domain

In the aggregation domain, IGP & LDP are not affected by the number of access nodes outside of their domain. They are not affected by the total number of AN nodes:

IGP:

$$\text{node} : \text{AGN} / \text{Area} \sim o(1)$$

$$\text{links} : 3 * \text{AGN} / \text{Area} \sim o(1)$$

$$\text{IP prefixes} : \text{Core} + \text{Area} + (5 * \text{AGN} + \text{AN}) / \text{Area} \sim o(\text{AN} * 5 / \text{Area})$$

+ + 1 loopback per core node + one aggregate per area + 5  
prefixes per AGN in the area + 1 prefix per AN in the area.

LDP FEC:

$$\text{Core} + (\text{AGN} + \text{AN}) / \text{Area} \sim o(\text{AN} / \text{Area})$$

+ + 1 loopback per core node + 1 loopback per AGN & AN node in  
the area.

AGN FIBs grows with the number of node in the core area, in their aggregation area, plus the number of inter domain LSP required by the AN attached to them. They do not depend on the total number of AN nodes. In the BGP control plane, AGN also needs to handle all the AN routes.

AGN:

IP FIB :  $\#Core + \#Area + (5 * \#AGN + \#AN) / \#Area \sim o(\#AN * 5 / \#Area)$

MPLS LFIB :  $\#Core + (\#AGN + \#AN) / \#Area + 100 \sim o(\#AN / \#Area)$

AN FIBs grows with its connectivity requirement. They do not depend on the number of AN, AGN, SN or any others nodes.

AN:

IP RIB :  $1 \sim o(1)$

MPLS LIB :  $1k \sim o(1)$

IP FIB :  $1 \sim o(1)$

MPLS LFIB :  $1k \sim o(1)$

#### 5.2.1.4. Summary

AN requirements are kept minimal. BGP is not required and the size of their FIB is limited to their own connectivity requirements.

In the core area, IGP and LDP are not affected by the node in the aggregation domains. In particular they do not grow with the number of AGN or AN.

In the aggregation areas, IGP and LDP are affected by the number of core nodes and the number of AGN and AN in their area. They are not affected by the total number of AGN or AN in the seamless MPLS domain.

No FIB of any node is required to handle the total number of AGN or AN in the seamless MPLS domain. In other word, the number of AGN and AN in the seamless MPLS domain is not limited, if the number of areas can grow accordingly. The main limitation is the MPLS connectivity requirements on the AN, i.e. mainly the number of LSP needed on the AN. Another limitation may be the number of different LSP needed by AN attached or behind an AGN. However, given foreseen deployments and current AGN capabilities, this is not expected to be a limitation.

In the control plane, BGP will typically handle all AN routes. This is significant but target deployments are well under current equipments capacities. In addition, if required, additional techniques could be used to improve this scalability, based on the experience gained with scaling BGP/MPLS VPN (e.g. route partitioning between RR planes, route filtering (static or dynamic with ORF or

route refresh) between AN and on AGN to improve AGN scalability.

#### 5.2.1.5. Numerical application for use case #1

As a recap, targets for deployment scenario 1 are:

- o Number of Aggregation Domains 100
- o Number of Backbone Nodes 1.000
- o Number of AGgregation Nodes 10.000
- o Number of Access Nodes 100.000

This gives the following scaling numbers for each category of nodes:

- o AN IP FIB 1
- o AN MPLS LFIB 1 000
- o AGN IP FIB 2 600
- o AGN MPLS LFIB 2 200
- o ABR IP FIB 7 600
- o ABR MPLS LFIB 2 100
- o TN IP FIB 5 000
- o TN MPLS LFIB 1 000
- o RR BGP NLRI 100 000
- o RR BGP paths 200 000

#### 5.2.1.6. Numerical application for use case #2

As a recap, targets for deployment scenario 1 are:

- o Number of Aggregation Domains 30
- o Number of Backbone Nodes 150
- o Number of AGgregation Nodes 1.500
- o Number of Access Nodes 40.000

This gives the following scaling numbers for each category of nodes:

- o AN IP FIB 1
- o AN MPLS LFIB 1 000
- o AGN IP FIB 1 700
- o AGN MPLS LFIB 1 800
- o ABR IP FIB 3 700
- o ABR MPLS LFIB 1 600
- o TN IP FIB 750
- o TN MPLS LFIB 150
- o RR BGP NLRI 40 000
- o RR BGP paths 80 000

## 6. Acknowledgements

Many people contributed to this document. The authors wish to thank - in alphabetical order:

- o Wim Henderickx (Alcatel)
- o Clarence Filsfils (Cisco Networks),
- o Thomas Beckhaus, Wilfried Maas, Roger Wenner (Deutsche Telekom),
- o Kireeti Kompella (Juniper Networks),

## 7. IANA Considerations

This memo includes no request to IANA.

All drafts are required to have an IANA considerations section (see the update of RFC 2434 [I-D.narten-iana-considerations-rfc2434bis] for a guide). If the draft does not require IANA to do anything, the section contains an explicit statement that this is the case (as above). If there are no requirements for IANA, the section will be removed during conversion into an RFC by the RFC Editor.

## 8. Security Considerations

In a typical MPLS deployment the use of MPLS is limited to relatively small network consisting of core and edge nodes. Those nodes are under full control of the services provider and placed at locations where only authorized personal has access (this also includes physical access to the nodes). With the extensions of MPLS towards access and aggregation nodes not all nodes will be "locked away" in secure locations. Small access nodes like DSLAMs will be located in street cabinets, potentially offering access to the "interested researcher". Nevertheless the unauthorized access to such in device SHOULD NOT impose any security risks to the MPLS infrastructure itself. Seamless MPLS must be stable regarding attacks against access and aggregation nodes running MPLS.

Levels of Security: tbd.

Access Network: tbd.

Aggregation Network: tbd.

Core Network: tbd.

## 9. References

### 9.1. Normative References

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

### 9.2. Informative References

- [ABRFRR] Rekhter, Y., "Local Protection for LSP tail-end node failure, MPLS World Congress 2009".
- [ACM01] "Archieving sub-second IGP convergence in large IP networks, ACM SIGCOMM Computer Communication Review, v.35 n.3", July 2005.
- [BGPPIC] "BGP PIC, Technical Report", November 2007.
- [I-D.filsfils-rtgwg-lfa-applicability] Filsfils, C., Francois, P., Shand, M., Decraene, B., Uttaro, J., Leymann, N., and M. Horneffer, "LFA applicability in SP networks", draft-filsfils-rtgwg-lfa-applicability-00 (work in progress), March 2010.

- [I-D.ietf-bfd-v4v6-lhop]  
Katz, D. and D. Ward, "BFD for IPv4 and IPv6 (Single Hop)", draft-ietf-bfd-v4v6-lhop-11 (work in progress), January 2010.
- [I-D.kothari-henderickx-l2vpn-vpls-multihomeing]  
Kothari, B., Kompella, K., Henderickx, W., and F. Balus, "BGP based Multi-homing in Virtual Private LAN Service", draft-kothari-henderickx-l2vpn-vpls-multihomeing-01 (work in progress), July 2009.
- [I-D.narten-iana-considerations-rfc2434bis]  
Narten, T. and H. Alvestrand, "Guidelines for Writing an IANA Considerations Section in RFCs", draft-narten-iana-considerations-rfc2434bis-09 (work in progress), March 2008.
- [I-D.raggarwa-mac-vpn]  
Aggarwal, R., Isaac, A., Uttaro, J., Henderickx, W., and F. Balus, "BGP MPLS Based MAC VPN", draft-raggarwa-mac-vpn-01 (work in progress), June 2010.
- [I-D.sajassi-l2vpn-rvpls-bgp]  
Sajassi, A., Patel, K., Mohapatra, P., Filsfils, C., and S. Boutros, "Routed VPLS using BGP", draft-sajassi-l2vpn-rvpls-bgp-01 (work in progress), July 2010.
- [PEFRR] Le Roux, J., Decraene, B., and Z. Ahmad, "Fast Reroute in MPLS L3VPN Networks - Towards CE-to-CE Protection, MPLS 2006 Conference".
- [RFC2629] Rose, M., "Writing I-Ds and RFCs using XML", RFC 2629, June 1999.
- [RFC3031] Rosen, E., Viswanathan, A., and R. Callon, "Multiprotocol Label Switching Architecture", RFC 3031, January 2001.
- [RFC3107] Rekhter, Y. and E. Rosen, "Carrying Label Information in BGP-4", RFC 3107, May 2001.
- [RFC3209] Awduche, D., Berger, L., Gan, D., Li, T., Srinivasan, V., and G. Swallow, "RSVP-TE: Extensions to RSVP for LSP Tunnels", RFC 3209, December 2001.
- [RFC3353] Ooms, D., Sales, B., Livens, W., Acharya, A., Griffoul, F., and F. Ansari, "Overview of IP Multicast in a Multi-Protocol Label Switching (MPLS) Environment", RFC 3353,

August 2002.

- [RFC3552] Rescorla, E. and B. Korver, "Guidelines for Writing RFC Text on Security Considerations", BCP 72, RFC 3552, July 2003.
- [RFC4090] Pan, P., Swallow, G., and A. Atlas, "Fast Reroute Extensions to RSVP-TE for LSP Tunnels", RFC 4090, May 2005.
- [RFC4364] Rosen, E. and Y. Rekhter, "BGP/MPLS IP Virtual Private Networks (VPNs)", RFC 4364, February 2006.
- [RFC5036] Andersson, L., Minei, I., and B. Thomas, "LDP Specification", RFC 5036, October 2007.
- [RFC5283] Decraene, B., Le Roux, J.L., and I. Minei, "LDP Extension for Inter-Area Label Switched Paths (LSPs)", RFC 5283, July 2008.
- [RFC5286] Atlas, A. and A. Zinin, "Basic Specification for IP Fast Reroute: Loop-Free Alternates", RFC 5286, September 2008.
- [RFC5332] Eckert, T., Rosen, E., Aggarwal, R., and Y. Rekhter, "MPLS Multicast Encapsulations", RFC 5332, August 2008.

#### Appendix A. ABR Fast Reroute

In order to deploy local repair also when an area border router fails and additional mechanism is used (see Figure 8) with the PLR being the point of local repair.

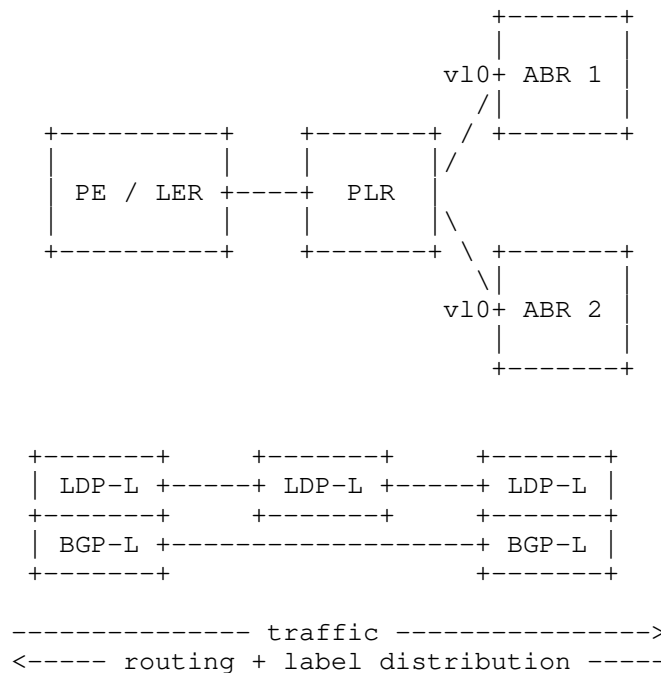


Figure 8: Routing and Traffic Flow

In order to deploy local repair also when an area border router fails an additional mechanism is used (see Figure 8) with the PLR being the point of local repair. ABR1 advertises in addition to it's loopback address in ISIS and LDP a virtual loopback address (vl0). ABR2 is backup router and advertises the same address as well with a worse metric. If ABR1 fails the PLR can immediately locally switch to the alternative path towards ABR2 (anycast behaviour). If this anycast address is used as next hop within the labeled BGP a local repair for the MPLS transport layer can be performed.

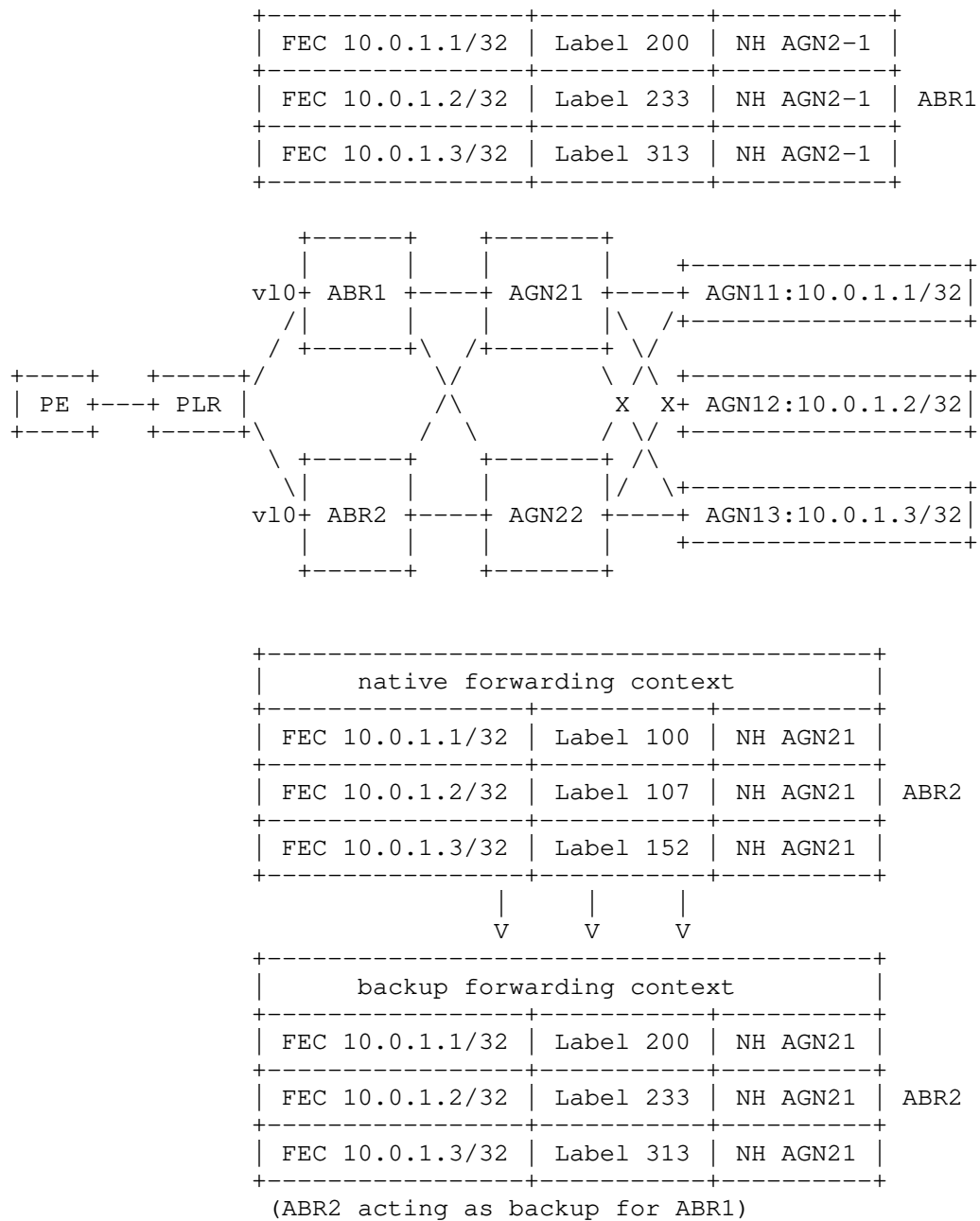


Figure 9: ABR Failure Scenario

Figure 9 shows the behaviour in case of an ABR failure. ABR2 creates

a backup forwarding context for labeled BGP routes which are routed via ABR1. This local backup forwarding context is not part of the global label table. ABR2 advertises over LDP as anycast next-hop not implicit null but an explicit label which points internally to the backup context. ABR2 learns via BGP the labels used by ABR1 for the connected aggregation domain(s). In addition ABR2 carries also label forwarding states for the corresponding FECs. By combining those two informations the backup forwarding context is filled with valid label forwarding entries. MPLS frames carrying the correct label for the forwarding by ABR1 as second entry in the label stack are also correctly forwarded by ABR2 using this information.

#### Authors' Addresses

Nicolai Leymann (editor)  
Deutsche Telekom AG  
Winterfeldtstrasse 21  
Berlin 10781  
DE

Phone: +49 30 8353-92761  
Email: n.leymann@telekom.de

Bruno Decraene  
France Telecom  
38-40 rue du General Leclerc  
Issy Moulineaux cedex 9, 92794  
FR

Phone:  
Fax:  
Email: bruno.decraene@orange-ftgroup.com  
URI:

Clarence Filsfils  
Cisco Systems  
Brussels,  
Belgium

Phone:  
Fax:  
Email: cfilsfil@cisco.com  
URI:

Dirk Steinberg  
Steinberg Consulting  
Ringstrasse 2  
Buchholz 53567  
DE

Email: dws@steinbergnet.net



Internet Engineering Task Force  
Internet-Draft  
Intended status: Standards Track  
Expires: April 21, 2011

L. Li  
L. Huang  
China Mobile, Inc.  
N. So  
Verison Business  
A. Kvalbein  
Resiliens Communication AS  
October 18, 2010

MPLS Multiple Topology Applicability and Requirements  
draft-li-mpls-mt-applicability-requirement-00.txt

#### Abstract

This document describes the applicability and requirements for Multiprotocol Label Switching Multiple Topology (MPLS-MT). The applicability and requirements are presented from different angles. They are expressed from a customer's point of view, a service provider's point of view and a vendor's point of view.

#### 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). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at <http://datatracker.ietf.org/drafts/current/>.

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

This Internet-Draft will expire on April 21, 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.

## Table of Contents

1. Terminology . . . . .	3
2. Introduction . . . . .	3
3. Applicability . . . . .	5
3.1. Simplified Data-plane . . . . .	5
3.2. Automation of inter-layer interworking . . . . .	6
3.3. Migration without service disruption . . . . .	6
3.4. Protection using MT . . . . .	7
3.5. Service Separation . . . . .	7
3.6. Load Balancing . . . . .	7
4. Requirements . . . . .	8
4.1. Service Requirements . . . . .	8
4.1.1. Availability . . . . .	8
4.1.2. Stability . . . . .	8
4.1.3. Traffic types . . . . .	8
4.1.4. Data isolation . . . . .	9
4.1.5. Security . . . . .	9
4.1.6. Topology . . . . .	10
4.1.7. Addressing . . . . .	10
4.1.8. Quality of Service . . . . .	11
4.1.9. Network Resource Partitioning and Sharing between MPLS-MTs (REWRITE with emphasis/focus on partition) . . . . .	11
4.2. Provider requirements . . . . .	11
4.2.1. Scalability . . . . .	11
4.2.2. Management . . . . .	13
4.2.3. Customer Management of a MPLS-MT . . . . .	14
4.3. Engineering requirements . . . . .	14
4.3.1. Forwarding plane requirements . . . . .	14
4.3.2. Control plane requirements . . . . .	15
4.3.3. Control Plane Containment . . . . .	15
4.3.4. Requirements for commonality of MPLS-MT mechanisms . .	15
4.3.5. Interoperability . . . . .	16
5. IANA Considerations . . . . .	16
6. Acknowledgement . . . . .	16
7. References . . . . .	16
7.1. Normative References . . . . .	16
7.2. Informative References . . . . .	17
Authors' Addresses . . . . .	17

## 1. Terminology

Terminology used in this document

Non-MT: router Routers that do not have the MT capability.

MT router: Routers that have MT capability as described in this document.

MT-ID: Renamed TOS field in LSAs to represent Multiple-TopologyID.

Default topology: Topology that is built using the TOS 0 metric (default metric).

MT topology: Topology that is built using the corresponding MT-ID metric.

MT: Shorthand notation for MPLS Multiple Topology.

MT#0 topology: Representation of TOS 0 metric in MT-ID format.

Non-MT-Area: An area that contains only non-MT routers.

MT-Area: An area that contains both non-MT routers and MT routers, or only MT routers.

## 2. Introduction

"Multi-Topology Routing in OSPF", RFC4915, describes a mechanism for Open Shortest Path First protocol to support Multi-Topologies (MTs) in IP network where the Type of Service (TOS) based metric fields are redefined and are used to advertise different topologies, each with a separate link metric. The classification of what type of traffic maps to which topology is not defined in RFC4915. The interface can be configured to belong to a set of topologies. Network topology changes will be advertised independently for each topology using a Multi-Topology Identifier (MT-ID), so that IP packets can be forwarded in the specific network topology independently.

"M-ISIS: Multi Topology (MT) Routing in Intermediate System to Intermediate Systems (IS-ISs)", RFC5120, describes a mechanism within Intermediate System to Intermediate Systems (IS-ISs) to run a set of independent IP topologies. The existing IS-IS protocol is extended so that the formation of adjacencies and advertising of prefixes and reachable intermediate system are performed independently within each topology.

There is a need to support Multiple-Topologies in MPLS network where a label switch path (LSP) is established within one topology or across multiple topologies and the traffic can be forwarded along the LSP within each network topology or across multiple topologies.

This document presents requirements for Multiprotocol Label Switching Multiple-Topology (MPLS-MT). It identifies requirements that may apply to one or more individual approaches that a Service Provider may use to provision LSPs in MPLS-MT. The specification of technical means to provide MPLS-MT services is outside the scope of this document. Other documents are intended to cover this aspect. This document is intended as a "checklist" of requirements, providing a consistent way to evaluate and document how well each approach satisfies specific requirements. The applicability statement documents for each approach should provide the results of this evaluation. This document is not intended to compare one approach to another. This document presents requirements from several points of view. It begins with some considerations from a point of view common to customers and service providers, continues with a customer perspective, and concludes with specific needs of a Service Provider (SP).

There are three different deployment scenarios MPLS-MT services are considered in this document:

1. Single-provider, single-AS: This is the least complex scenario, where the MPLS-MT service is offered across a single service provider network spanning a single Autonomous System.
2. Single-provider, multi-AS: In this scenario, a single provider may have multiple Autonomous Systems (e.g., a global Tier-1 ISP with different ASes depending on the global location, or an ISP that has been created by mergers and acquisitions of multiple networks). This scenario involves the constrained distribution of routing information across multiple Autonomous Systems.
3. Multi-provider: This scenario is the most complex, wherein trust negotiations need to be made across multiple service provider backbones in order to meet the security and service level agreements for the MPLS-MT customer. This scenario can be generalized to cover the Internet, which comprises of multiple service provider networks. It should be noted that customers can construct their own MPLS-MTs across multiple providers. However such MPLS-MTs are not considered here as they would not be "Providerprovisioned".

MPLS-MT is set of extensions to existing MPLS signaling protocols that makes MPLS signaling protocols aware of multi-topology. In the context of MPLS signaling the term "Multi-topology" is redefined to

be protocol independent unlike IGP-MT, which is scoped inside a single flavor IGP (ex. ISIS-MT or OSPF-MT). In other words, a MPLS Multiple-Topology can be mapped to a OSPFv2 based topology, OSPFv3 based topology or ISIS based IGP topology. Besides, a MPLS multi-topology can also be mapped to an instance of OSPF-MT or ISIS-MT. There are two major categories for MPLS-MT applications: a) MPLS RSVP-TE-MT applications, b) MPLS LDP-MT applications. The following sections of the draft describe application scenarios and MPLS-MT signaling in general. These application scenarios are useful for service providers who already have an MPLS network, or for service providers willing to migrate from IP to MPLS.

The following Sections describe applicability and generic MPLS-MT requirements.

### 3. Applicability

There are two main scenarios for how MPLS-MT can be used as a value-adding tool: 1) It can be exposed to and used by the customer to suit particular needs. For example, a customer might be given the option to select from a range of different topologies with different price and quality characteristics, and can select one (or more) that fulfils the given requirements. This could allow a service provider to better exploit network resources, by using pricing as an incentive. 2) It can be used as a management tool by the network operator to achieve certain goals such as resilience, traffic isolation and congestion avoidance, without exposing this to customers. Of course, one scenario does not exclude the other: an operator might want to offer MT routing to large customers, while also using it as a tool for "internal" purposes for its best effort services.

#### 3.1. Simplified Data-plane

IGP-MT requires additional data-plane resources to maintain a separate forwarding table for each configured MT. On the other hand, MPLS-MT does not change the data-plane system architecture, if an IGP-MT is mapped to an MPLS-MT. In case MPLS-MT, the incoming label value itself can determine an MT, and hence it requires a single NHLFE space. MPLS-MT requires only MT-RIBs in the control-plane, and there is no need to have extra MT-FIBs. Forwarding IP packets over a particular MT requires either configuration or some external means at every node, to map an attribute of incoming IP packet header to IGP-MT, which is additional overhead for network management. With MPLS-MT, mapping is required only at the ingress-PE of an MPLS-MT LSP, because each node identifies MPLS-MT LSP switching based on incoming label, hence no additional configuration is required at

every node.

### 3.2. Automation of inter-layer interworking

With (G)MPLS-RSVP-MT extensions, an ingress-PE can signal a particular path (ERO) that can traverse different network layers to reach a egress-PE. For instance, an ERO is associated with MT-ID RSVP subobject to indicate a "P" router to use a particular Layer-1 TE- link-state topology, instead of the default Layer-3 link-state topology as illustrated in the following diagram. With this mechanism a (G)MPLS-TE LSP can be offloaded to lower layers without service disruption and without complexity of configuration.

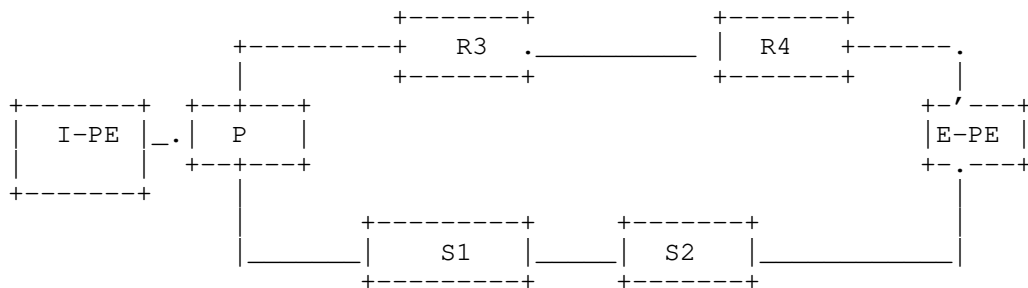


Figure 1: Layer-3 Link State Topology

Layer-3 ERO : P[MT-0]->R3->R4->E-PE[MT-0].

Inter-layer ERO : P[MT-0]->loose-hop[MT-1]->E-PE[MT-0]

Procedures to discover MT mapping with an IGP topology at ingress-PE nodes requires some auto-discovery mechanism.

Figure 1: Layer-3 Link State Topology

### 3.3. Migration without service disruption

As stated above, MPLS-MT abstracts link state topology and identifies it by a unique MT-ID, which need not be the same as the IGP-MT ID. This characteristic is quite useful for service providers looking to migrate to a different flavor of IGP, e.g., OSPFv2 to ISIS6, OSPFv2 to OSPFv3. Service providers would like to incrementally upgrade their topologies, which requires an LSP to traverse multiple IGP

domains (OSPFv2 to OSPFv3) or (OSPF to ISIS). Migrating TE-LSPs to use a newly deployed link state topology requires a non-trivial effort. This migration may involve service disruption, especially when a path includes loose-hops in the ERO. For example: When an incoming PATH message requires an LSR to resolve loose-hop over a newly deployed IGP domain, which is not possible in the absence of MPLS-MT signaling. MPLS-MT allows an ingress-PE to specify Multiple-Topology to be used at every hop.

### 3.4. Protection using MT

We know that [IP-FRR-MT] can be used for configuring alternate paths via backup-mt, such that if the primary link fails, then a backup-MT can be used for forwarding. However, such techniques require special marking of IP packets that are forwarded using backup-MTs. MPLS-LDP-MT procedures simplify the forwarding of the MPLS packets over backup-MTs, as the MPLS-LDP-MT procedure distributes separate labels for each MT. How backup paths are computed depends on the implementation, and the algorithm. MPLS-RSVP-MT in conjunction with IGP-MT could be used to separate the primary traffic and backup traffic. For example, service providers can create a backup MT that consists of links that are meant only for backup traffic. Service providers can then establish bypass LSPs, standby LSPs, using backup MT, thus keeping undeterministic backup traffic away from the primary traffic.

### 3.5. Service Separation

MPLS-MT procedures allow establishing two distinct LSPs for the same FEC, by advertising a separate label mapping for each configured topology. Service providers can implement CoS using MPLS-MT procedures without requiring to create a separate FEC address for each class. MPLS-MT can also be used to separate multicast and unicast traffic.

### 3.6. Load Balancing

MPLS-MT can be used to construct several alternative LSPs between PE routers. The LSPs in different topologies might follow partly overlapping routes through the network, or be completely disjoint. By smart assignment traffic to different MTs at the PE routers, it is possible to offload traffic from heavily loaded links, and hence reduce the risk of congestion and improve resource utilization. This type of load balancing can be performed either in an offline way, where traffic is assigned to each MT according to a static split ratio, or in an online fashion, where the amount of traffic assigned to each MT according to a dynamic splitting function that depends on the current load situation.

## 4. Requirements

### 4.1. Service Requirements

These are the requirements that a customer can observe or measure for verifying whether the MPLS-MT service that the Service Provider (SP) provides is satisfactory. As mentioned before, each of these requirements apply equally across each of the three deployment scenarios unless stated otherwise.

#### 4.1.1. Availability

MPLS-MT services **MUST** have high availability. LSPs that cross over several MTs require connectivity to be maintained even in the event of network failures.

This can be achieved via various redundancy techniques such as:

##### 4.1.1.1. Physical Diversity and FRR

A single MT router may be connected to multiple MT routers. For a LSP, both local protections and global protections can be set up. Thus when a network failure happens, the traffic carried by the LSP can continue to flow across the MTs from the head end of the LSP to the tail ends of the LSP.

It should be noted that it is difficult to guarantee high availability when the MPLS-MT service is across multiple providers, unless there is a negotiation between the different service providers to maintain the service level agreement for the MPLS-MT customer.

#### 4.1.2. Stability

In addition to availability, MPLS-MT services **MUST** also be stable. Stability is a function of several components such as MT routing and MPLS-MT signaling. For example, in the case of MT routing, route flapping or routing loops **MUST** be avoided in order to ensure stability. Stability of the MPLS-MT service is directly related to the stability of the mechanisms and protocols used to establish LSPs. It **SHOULD** also be possible to allow network upgrades and maintenance procedures without impacting the MPLS-MT service.

#### 4.1.3. Traffic types

MPLS-MT services **MUST** support unicast (or point to point) traffic and **SHOULD** support multicast (or point-to-multipoint) traffic. For multicast traffic, the network delivers a stream to a set of destinations that have registered interest in the stream through a

P2MP LSP. It is desirable to support multicast limited in scope to an intranet or extranet. The solution SHOULD be able to support a large number of such intranet or extranet specific multicast groups in a scalable manner. All MPLS-MT approaches SHALL support both IPv4 and IPv6 traffic.

#### 4.1.4. Data isolation

The MPLS-MT MUST support forwarding plane isolation. The network MUST never deliver user data across MPLS-MT boundaries unless the two MPLS-MTs participate in an intranet or extranet.

Furthermore, if the provider network receives signaling or routing information from one MPLS-MT, it MUST NOT reveal that information to another MPLS-MT unless the two MPLS-MTs participate in an intranet or extranet. It should be noted that the disclosure of any signaling/routing information across an extranet MUST be filtered per the extranet agreement between the organizations participating in the extranet.

#### 4.1.5. Security

A range of security features SHOULD be supported by the suite of MPLS-MT solutions in the form of securing customer flows, providing authentication services for temporary, remote or mobile users, and the need to protect service provider resources involved in supporting a MPLS-MT. Each MPLS-MT solution SHOULD state which security features it supports and how such features can be configured on a per customer basis. Protection against Denial of Service (DoS) attacks is a key component of security mechanisms.

Some security mechanisms may be equally useful regardless of the scope of the MPLS-MT. Other mechanisms may be more applicable in some scopes than in others. For example, in some cases of single-provider single-AS MPLS-MTs, the MPLS-MT service may be isolated from some forms of attack by isolating the infrastructure used for supporting MPLS-MTs from the infrastructure used for other services. However, the requirements for security are common regardless of the scope of the MPLS-MT service.

##### 4.1.5.1. User data security

MPLS-MT solutions that support user data security SHOULD use standard methods to achieve confidentiality, integrity, authentication and replay attack prevention. Such security methods MUST be configurable between different end points. It is also desirable to configure security on a per-LSP basis. User data security using encryption is especially desirable in the multi-provider scenario.

#### 4.1.5.2. Access control

A MPLS-MT solution may also have the ability to activate the appropriate filtering capabilities upon request of a customer. A filter provides a mechanism so that access control can be invoked at the point(s) of communication between different organizations involved in an extranet. Access control can be implemented by a firewall, access control lists on routers, cryptographic mechanisms or similar mechanisms to apply policy-based access control. Such access control mechanisms are desirable in the multi-provider scenario.

#### 4.1.5.3. MT router authentication and authorization

A MPLS-MT solution requires authentication and authorization of the following:

1. temporary and permanent access for users connecting to a MT router (authentication and authorization BY the MT router)
2. the MT router itself (authentication and authorization FOR the MT router)

#### 4.1.5.4. Inter domain security

The MPLS-MT solution MUST have appropriate security mechanisms to prevent the different kinds of Distributed Denial of Service (DDoS) attacks, misconfiguration or unauthorized accesses in inter domain MPLS-MT connections. This is particularly important for multiservice provider deployment scenarios. However, this will also be important in single-provider multi-AS scenarios.

#### 4.1.6. Topology

An MPLS-MT implementation SHOULD support arbitrary, customer -defined connectivity to the extent possible, for example, from partial mesh to full mesh topology. These can actually be different from the topology used by the service provider. The MPLS-MT services SHOULD be independent of MPLS-MT technology. To the extent possible, a MPLS-MT service SHOULD be independent of the geographic extent of the deployment. Multiple MPLS-MTs per customer SHOULD be supported without requiring additional hardware resources.

#### 4.1.7. Addressing

Each customer resource MUST be identified by an address that is unique within its MPLS-MT. It need not be identified by a globally unique address. Support for private addresses as described in

[RFC1918], as well as overlapping customer addresses SHALL be supported. One or more MPLS-MTs for each customer can be built over the same infrastructure without requiring any of them to renumber. The solution MUST NOT use NAT on the customer traffic to achieve that goal. Interconnection of two networks with overlapping IP addresses is outside the scope of this document.

#### 4.1.8. Quality of Service

A technical approach for supporting MPLS-MTs SHALL be able to support QoS via IETF standardized mechanisms such as Diffserv. Support for best-effort traffic SHALL be mandatory for all MPLS-MT types. The extent to which any specific MPLS-MT service will support QoS is up to the service provider. In many cases single-provider single-AS MPLS-MTs will offer QoS guarantees. Support of QoS guarantees in the multiservice-provider case will require cooperation between the various service providers involved in offering the service.

#### 4.1.9. Network Resource Partitioning and Sharing between MPLS-MTs (REWRITE with emphasis/focus on partition)

Network resources such as memory space, FIB table, bandwidth and CPU processing SHALL be shared between MPLS-MTs and, where applicable, with non-MPLS-MT Internet traffic. Mechanisms SHOULD be provided to prevent any specific MPLS-MT from taking up available network resources and causing others to fail. SLAs to this effect SHOULD be provided to the customer. Similarly, resources used for control plane mechanisms are also shared. When the service provider's control plane is used to distribute MPLS-MT specific information and provide other control mechanisms for MPLS-MTs, there SHALL be mechanisms to ensure that control plane performance is not degraded below acceptable limits when scaling the MPLS-MT service, or during network events such as failure, routing instabilities etc. Since a service provider's network would also be used to provide Internet service, in addition to MPLS-MTs, mechanisms to ensure the stable operation of Internet services and other MPLS-MTs SHALL be made in order to avoid adverse effects of resource hogging by large MPLS-MT customers.

#### 4.2. Provider requirements

This section describes operational requirements for a cost-effective, profitable MPLS-MT service offering.

##### 4.2.1. Scalability

The scalability for MPLS-MT solutions has many aspects. The list below is intended to comprise of the aspects that MPLS-MT solutions

SHOULD address. Clearly these aspects in absolute figures are very different for different types of MPLS -MTs. It is also important to verify that MPLS-MT solutions not only scales on the high end, but also on the low end - i.e., a MPLS-MT with three nodes and three users should be as viable as a MPLS-MT with hundreds of nodes and thousands of users.

#### 4.2.1.1. Service Provider Capacity Sizing Projections

A MPLS-MT solution SHOULD be scalable to support a large number of MPLS-MTs per Service Provider network.

A MPLS-MT solution SHOULD be scalable to support of a large number of routes per MPLS-MT. The number of routes per MPLS-MT may range from just a few to ( $O(10^5)$ ) exchanged between ISPs, with typical values being in the  $O(10^3)$  range. The high end number is especially true considering the fact that many large ISPs may provide MPLS-MT services to smaller ISPs or large corporations.

A MPLS-MT solution SHOULD support high values of the frequency of configuration setup and change. Approaches SHOULD articulate scaling and performance limits for more complex deployment scenarios, such as single-provider multi-AS MPLS-MTs, multi-provider MPLS-MTs. Approaches SHOULD also describe other dimensions of interest, such as capacity requirements or limits, number of interworking instances supported as well as any scalability implications on management systems. A MPLS-MT solution SHOULD support a large number of customer interfaces on a single PE or CE with current Internet protocols.

#### 4.2.1.2. MPLS-MT Scalability aspects

This section describes the metrics for scaling MPLS-MT solutions. These numbers are only representative and different service providers may have different requirements for scaling. Further discussion on service provider sizing projections is in Section 5.1.1. It should also be noted that the numbers given below would be different depending on whether the scope of the MPLS-MT is single-provider single-AS, single-provider multi-AS, or multiprovider. Clearly, the larger the scope, the larger the numbers that may need to be supported. However, this also means more management issues. The numbers below may be treated as representative of the single-provider case.

#### 4.2.1.3. Number of MPLS-MTs in the network

The number of MPLS-MTs SHOULD scale linearly with the size of the access network and with the number of PEs. The number of MPLS-MTs in

the network SHOULD be  $O(10)$ . This requirement also effectively places a requirement on the number of tunnels that SHOULD be supported in the network.

#### 4.2.1.4. Number of MPLS-MTs per customer

In some cases a service provider may support multiple MPLS-MTs for the same customer of that service provider. For example, this may occur due to differences in services offered per MPLS-MT (e.g., different QoS, security levels, or reachability) as well as due to the presence of multiple workgroups per customer. It is possible that one customer will run up to  $O(10)$  MPLS-MTs.

#### 4.2.1.5. Number of addresses and address prefixes per MPLS-MT

Since any MPLS-MT solution SHALL support private customer addresses, the number of addresses and address prefixes are important in evaluating the scaling requirements. The number of address prefixes used in routing protocols and in forwarding tables specific to the MPLS-MT needs to scale from very few (for smaller customers) to very large numbers seen in typical Service Provider backbones. The high end is especially true considering that many Tier 1 SPs may provide MPLS-MT services to Tier 2 SPs or to large corporations. This number would be on the order of addresses supported in typical native backbones.

#### 4.2.1.6. Solution-Specific Metrics

Each MPLS-MT solution SHALL document its scalability characteristics in quantitative terms. A MPLS-MT solution SHOULD quantify the amount of state that a PE and P device has to support. This SHOULD be stated in terms of the order of magnitude of the number of MPLS-MTs supported by the service provider.

#### 4.2.2. Management

A service provider MUST have a means to view the topology, operational state, service order status, and other parameters associated with each customer's MPLS-MT. Furthermore, the service provider MUST have a means to view the underlying logical and physical topology, operational state, provisioning status, and other parameters associated with the equipment providing the MPLS -MT service(s) to its customers.

In the multi-provider scenario, it is unlikely that participating providers would provide each other a view to the network topology and other parameters mentioned above. However, each provider MUST ensure via management of their own networks that the overall MPLS -MT

service offered to the customers are properly managed. In general the support of a single MPLS-MT spanning multiple service providers requires close cooperation between the service providers. One aspect of this cooperation involves agreement on what information about the MPLS-MT will be visible across providers, and what network management protocols will be used between providers. MPLS-MT devices SHOULD provide standards-based management interfaces wherever feasible.

#### 4.2.3. Customer Management of a MPLS-MT

A customer SHOULD have a means to view the topology, operational state, service order status, and other parameters associated with his or her MPLS-MT.

A customer SHOULD be able to make dynamic requests for changes to traffic parameters. A customer SHOULD be able to receive real-time response from the SP network in response to these requests. One example of such service is a "Dynamic Bandwidth management" capability, that enables real-time response to customer requests for changes of allocated bandwidth allocated to their MPLS-MT(s). A possible outcome of giving customers such capabilities is Denial of Service attacks on other MPLS-MT customers or Internet users. This possibility is documented in the Security Considerations section.

#### 4.3. Engineering requirements

These requirements are driven by implementation characteristics that make service and provider requirements achievable.

##### 4.3.1. Forwarding plane requirements

The SP is REQUIRED to provide per-MPLS-MT management, tunnel maintenance and other maintenance required in order to meet the SLA/SLS.

By definition, MPLS-MT traffic SHOULD be segregated from each other, and from non-MPLS-MT traffic in the network. After all, MPLS-MTs are a means of dividing a physical network into several logical or physical networks. MPLS-MT traffic separation SHOULD be done in a scalable fashion. However, safeguards SHOULD be made available against misbehaving MPLS-MTs to not affect the network and other MPLS-MTs.

A MPLS-MT solution SHOULD NOT impose any hard limit on the number of MPLS-MTs provided in the network.

#### 4.3.2. Control plane requirements

The plug and play feature of a MPLS-MT solution with minimum configuration requirements is an important consideration. The MPLS-MT solutions SHOULD have mechanisms for protection against customer interface and/or routing instabilities so that they do not impact other customers' services or impact general Internet traffic handling in any way.

A MPLS-MT SHOULD be provisioned with minimum number of steps. For this to be accomplished, an auto-configuration and an auto-discovery protocol, which SHOULD be as common as possible to all MPLS-MT solutions, SHOULD be defined. However, these mechanisms SHOULD NOT adversely affect the cost, scalability or stability of a service by being overly complex, or by increasing layers in the protocol stack.

Mechanisms to protect the SP network from effects of misconfiguration of MPLS-MTs SHOULD be provided. This is especially of importance in the multi-provider case, where misconfiguration could possibly impact more than one network.

#### 4.3.3. Control Plane Containment

The MPLS-MT control plane MUST include a mechanism through which the service provider can filter MPLS-MT related control plane information as it passes between Autonomous Systems. For example, if a service provider supports a MPLS-MT offering, but the service provider's neighbors do not participate in that offering, the service provider SHOULD NOT leak MPLS-MT control information into neighboring networks. Neighboring networks MUST be equipped with mechanisms that filter this information should the service provider leak it. This is important in the case of multi-provider MPLS-MTs as well as singleprovider multi-AS MPLS-MTs.

#### 4.3.4. Requirements for commonality of MPLS-MT mechanisms

The mechanisms used to establish a MPLS-MT service SHOULD re-use well-known IETF protocols as much as possible. It should, however, be noted that the use of Internet mechanisms for the establishment and running of an Internet-based MPLS-MT service, SHALL NOT affect the stability, robustness, and scalability of the Internet or Internet services. In other words, these mechanisms SHOULD NOT conflict with the architectural principles of the Internet, nor SHOULD it put at risk the existing Internet systems.

In addition to commonality with generic Internet mechanisms, infrastructure mechanisms used in different MPLS-MT solutions SHOULD be as common as possible.

#### 4.3.5. Interoperability

Each technical solution is expected to be based on interoperable Internet standards.

Multi-vendor interoperability at network element, network and service levels among different implementations of the same technical solution SHOULD be ensured (that will likely rely on the completeness of the corresponding standard). This is a central requirement for SPs and customers.

The technical solution MUST be multi-vendor interoperable not only within the SP network infrastructure, but also with the customer's network equipment and services making usage of the MPLS-MT service.

Inter-domain interoperability - It SHOULD be possible to deploy a MPLS-MT solution across domains, Autonomous Systems, or the Internet.

#### 5. IANA Considerations

TBD

#### 6. Acknowledgement

Thanks for the contributions from Quintin Zhao, Ravi Tori, Huaimo Chen, Luyuang Fang, Chao Zhou.

#### 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.
- [RFC3692] Narten, T., "Assigning Experimental and Testing Numbers Considered Useful", BCP 82, RFC 3692, January 2004.
- [RFC4915] Psenak, P., Mirtorabi, S., Roy, A., Nguyen, L., and P. Pillay-Esnault, "Multi-Topology (MT) Routing in OSPF", RFC 4915, June 2007.
- [RFC5120] Przygienda, T., Shen, N., and N. Sheth, "M-ISIS: Multi Topology (MT) Routing in Intermediate System to Intermediate Systems (IS-IS)", RFC 5120, February 2008.

- [RFC3209] Awduche, D., Berger, L., Gan, D., Li, T., Srinivasan, V., and G. Swallow, "RSVP-TE: Extensions to RSVP for LSP Tunnels", RFC 3209, December 2001.
- [RFC4420] Farrel, A., Papadimitriou, D., Vasseur, J., and A. Ayyangar, "Encoding of Attributes for Multiprotocol Label Switching (MPLS) Label Switched Path (LSP) Establishment Using Resource ReserVation Protocol-Traffic Engineering (RSVP-TE)", RFC 4420, February 2006.

## 7.2. Informative References

### Authors' Addresses

Lianyuan Li  
China Mobile, Inc.  
53A, Xibianmennei Ave.  
Xunwu District, Beijing 01719  
China  
  
Email: lilianyuan@chinamobile.com

Lu Huang  
China Mobile, Inc.  
53A, Xibianmennei Ave.  
Xunwu District, Beijing 01719  
China  
  
Email: huanglu@chinamobile.com

Ning So  
Verison Business  
2400 North Glenville Drive  
Richardson, TX 75082  
USA  
  
Email: Ning.So@verizonbusiness.com

Amund Kvalbein  
Resiliens Communication AS  
Martin Linges v 17, Fornebu  
Fornebu, Lysaker 1325  
Norway

Email: Amundk@simula.com

Quintin Zhao  
Huawei Technology  
125 Nagog Park  
Acton, MA 01919  
US

Phone:  
Email: qzhao@huawei.com

Huaimo Chen  
Huawei Technology  
125 Nagog Park  
Acton, MA 01919  
US

Phone:  
Email: huaimochen@huawei.com

Ravi Tori  
Juniper Networks

pratiravi@gmail.com



Network Working Group  
Internet Draft  
Intended Status: Standards Track  
Expiration Date: January 7, 2011

Kamran Raza  
Cisco Systems  
  
Sami Boutros  
Cisco Systems

July 8, 2010

LDP Typed Wildcard PW FEC Elements  
draft-raza-l2vpn-pw-typed-wc-fec-01.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 January 7, 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 BSD License.

## Abstract

An extension to the Label Distribution Protocol (LDP) defines the general notion of a "Typed Wildcard Forwarding Equivalence Class (FEC) Element". This can be used when it is desired to request all label bindings for a given type of FEC Element, or to release or withdraw all label bindings for a given type of FEC element. However, a typed wildcard FEC element must be individually defined for each type of FEC element. This specification defines the typed wildcard FEC elements for the Pseudowire Identifier (PW Id) and Generalized Pseudowire Identifier (Gen. PW Id) FEC types.

## Conventions used in 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 [RFC2119].

## Table of Contents

1. Introduction	3
2. Typed Wildcard for PWid FEC Element	3
3. Typed Wildcard for Generalized PWid FEC Element	3
4. Operation	3
4.1. PW Consistency Check	4
4.2. PW Graceful Shutdown	5
5. Security Considerations	5
6. IANA Considerations	5
7. Acknowledgments	5
8. References	5
8.1. Normative References	5
8.2. Informative References	6
Author's Address	6

## 1. Introduction

An extension [TYPED-WC] to the Label Distribution Protocol (LDP) [RFC5036] defines the general notion of a "Typed Wildcard Forwarding Equivalence Class (FEC) Element". This can be used when it is desired to request all label bindings for a given type of FEC Element, or to release or withdraw all label bindings for a given type of FEC element. However, a typed wildcard FEC element must be individually defined for each type of FEC element.

[RFC4447] defines the "Pwid FEC Element" and "Generalized Pwid FEC Element" but it does not specify Typed Wildcard format for these elements. This document specifies the format of the Typed Wildcard FEC for the "Pwid FEC Element" and the "Generalized Pwid FEC Element" defined in [RFC4447]. The procedures for Typed Wildcard processing for Pwid and Generalized Pwid FEC Elements are same as described in [TYPED-WC] for any typed wildcard FEC Element type.

## 2. Typed Wildcard for Pwid FEC Element

The format of the Pwid FEC Typed Wildcard FEC is:

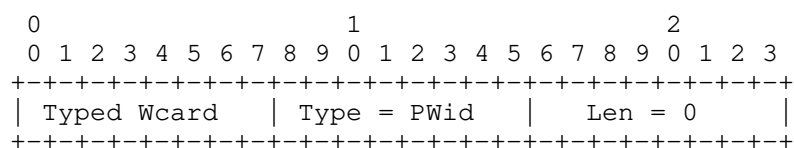


Figure 1: Format of Pwid Typed Wildcard FEC Element

Where:

Typed Wcard (one octet): as specified in [TYPED-WC]

FEC Element Type (one octet): Pwid FEC Element (type 0x80 [RFC4447])

Len FEC Type Info (one octet): Zero. (There is no additional FEC info)

## 3. Typed Wildcard for Generalized Pwid FEC Element

The format of the Generalized Pwid FEC Typed Wildcard FEC is:

```

      0                               1                               2
      0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
| Typed Wcard | Type=Gen.PWid | Len = 0 |
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+

```

Figure 2: Format of Generalized PWid Typed Wildcard FEC Element

Where:

Typed Wcard (one octet): as specified in [TYPED-WC]

FEC Element Type (one octet): Generalized PWid FEC Element (type 0x81 [RFC4447])

Len FEC Type Info (one octet): Zero. (There is no additional FEC info)

When Generalized PWid FEC Typed Wildcard is used, "PW Grouping ID TLV" [RFC4447] MUST NOT be present in the same message.

#### 4. Operation

The use of Typed Wildcard FEC elements for PW can be useful under several scenarios. This section describes two use cases to illustrate their usage. The following use cases consider two LSR nodes, A and B, with LDP session between them to exchange L2VPN PW bindings.

##### 4.1. PW Consistency Check

A user may request a control plane consistency check at LSR A for the PWid FEC and Generalized PWid FEC bindings that it had learnt from LSR B over LDP session. To perform this consistency check, LSR A marks all its learnt PW bindings from LSR B as stale, and then sends a Label Request message towards LSR B with Typed Wildcard FEC element for PWid FEC element and Generalized PWid FEC element. Upon receipt of such request, LSR B replays its database related to PWid FEC elements and Generalized PWid FEC element in Label Mapping message. As a PW binding is received at LSR A, the associated binding state is marked as refreshed (no stale). When replay completes for a given type of FEC, LSR B sends End-of-LIB Notification [END-OF-LIB] to mark the end of update for the given FEC type. Upon receipt of this Notification at LSR A, any remaining stale PW binding of given FEC type learnt from the peer LSR B, is

cleaned up and removed from the database. This completes consistency check with LSR B at LSR A for given FEC type.

#### 4.2. PW Graceful Shutdown

It may be desirable to perform shutdown/removal of existing PW bindings advertised towards a peer in a graceful manner -

- i.e. all

advertised PW bindings to be removed from a peer without session flap. For example, to request a graceful delete of the PWid FEC and Generalized PWid FEC bindings at LSR A learnt from LSR B, LSR A would send a Label Withdraw message towards LSR B with Typed Wildcard FEC elements pertaining to PWid FEC element and Generalized PWid FEC element. Upon receipt of such message, LSR B will delete all PWid and Generalized PWid bindings learnt from LSR A. Afterwards, LSR B would send Label Release message corresponding to received Label Withdraw with Typed FEC element.

#### 5. Security Considerations

No new security considerations beyond that apply to the base LDP specification [RFC5036], [RFC4447] and [MPLS\_SEC] apply to the use of the PW Typed Wildcard FEC Element types described in this document.

#### 6. IANA Considerations

This document defines no new element for IANA Consideration.

#### 7. Acknowledgments

The authors would like to thank Eric Rosen, M. Siva, and Zafar Ali for their valuable comments.

This document was prepared using 2-Word-v2.0 template.dot.

#### 8. References

##### 8.1. Normative References

[RFC5036] Andersson, L., Menei, I., and Thomas, B., Editors, "LDP Specification", RFC 5036, September 2007.

[TYPED-WC] Thomas, B., Asati, R., and Minei, I., "LDP Typed Wildcard FEC", draft-ietf-mpls-ldp-typed-wildcard-07.txt, Work in Progress, March 2010.

[END-OF-LIB] Asati, R., Mohapatra, P., Chen, E., and Thomas, B., "Signaling LDP Label Advertisement Completion", draft-ietf-mpls-ldp-end-of-lib-04.txt, Work in Progress, June 2010.

[RFC4447] L. Martini, Editor, E. Rosen, El-Aawar, T. Smith, G. Heron, "Pseudowire Setup and Maintenance using the Label Distribution Protocol", RFC 4447, April 2006.

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

## 8.2. Informative References

[MPLS\_SEC] Fang, L. et al., "Security Framework for MPLS and GMPLS Networks", draft-ietf-mpls-mpls-and-gmpls-security-framework-05.txt, Work in Progress, March 2009.

## Author's Address

Syed Kamran Raza  
Cisco Systems, Inc.,  
2000 Innovation Drive,  
Kanata, ON K2K-3E8, Canada.  
E-mail: skraza@cisco.com

Sami Boutros  
Cisco Systems, Inc.  
3750 Cisco Way,  
San Jose, CA 95134, USA.  
E-mail: sboutros@cisco.com

Network Working Group  
Internet Draft  
Intended status: Standards Track  
Expires: April 17, 2011

Kamran Raza  
Cisco Systems

Sami Boutros  
Cisco Systems

October 18, 2010

## LDP IP and PW Capability

draft-raza-mpls-ldp-ip-pw-capability-00.txt

### 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), 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 April 17, 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.

## Abstract

Currently, no LDP capability is exchanged for LDP applications like IP label switching and L2VPN/PW signaling. When an LDP session comes up, an LDP speaker may unnecessarily advertise its local state for such LDP applications even when the peer session may be established for some other applications like ICCP. This document proposes a solution by which an LDP speaker announces its "incapability" or disability or non-support for IP label switching or L2VPN/PW application, hence disabling corresponding application state exchange over established LDP session.

## Table of Contents

1. Introduction	3
2. Conventions used in this document	3
3. Non-negotiated LDP applications	4
3.1. Application Control Capabilities	4
3.1.1. IP Label Switching Capability TLV	4
3.1.2. PW Signaling Capability TLV	5
3.2. Procedures for Application Control Capabilities in an Initialization message	6
3.3. Procedures for Application Control capabilities in a Capability message	7
4. Operational Examples	8
4.1. Disabling IP/PW label applications on an ICCP session	8
4.2. Disabling IP Label Switching application on a PW session	8
4.3. Disabling IP application dynamically on an established IP/PW session	9
5. Security Considerations	9
6. IANA Considerations	9
7. Conclusions	10
8. References	10
8.1. Normative References	10
8.2. Informative References	10
9. Acknowledgments	10

## 1. Introduction

LDP Capabilities [RFC5561] introduced a mechanism to negotiate LDP capabilities for given feature amongst peer LSRs. This mechanism insures that no unnecessary state is exchanged between peer LSRs unless corresponding feature capability is successfully negotiated between peers.

While new features and applications like Typed Wildcard FEC [RFC5918], Inter-Chassis Communication Protocol [ICCP], and mLDP [MLDP] make use of LDP capabilities framework for their feature negotiation, the earlier LDP features and applications like IP label switching and L2VPN/PW signaling [RFC4447] may cause unnecessary state exchange between LDP peers if the given application is not enabled on one of the LDP speakers participating in a given session. For example, when bringing up and using an LDP peer session with a remote PE LSR for purely ICCP signaling purposes, the LDP speaker may unnecessarily advertise labels for IP (unicast) prefixes to this ICCP related LDP peer as per its default behavior. To avoid this unnecessary state advertisement and exchange, currently customers are typically required to configure/define some sort of LDP state/label filtering policies on the box, which introduces operational overhead and complexity.

This document proposes a solution by which an LDP speaker announces its "incapability" (or disability) to its peer for IP Label Switching and/or L2VPN/PW Signaling application at session establishment time. This helps avoiding unnecessary state exchange for such feature applications. The proposal also allows a previously disabled application to be enabled later during the session lifetime. The document introduces two new LDP Capabilities for IP label switching and L2VPN/PW applications to implement the proposal.

## 2. Conventions used in 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 RFC-2119 [RFC2119].

The term "IP" in this document refers to "IP unicast", unless otherwise explicitly stated.

### 3. Non-negotiated LDP applications

For the applications that existed before LDP Capabilities [RFC5561] mechanics were defined, LDP speaker may advertise relevant application state to its peers after session establishment without waiting for any capabilities exchange and negotiation.

The most important non-negotiated applications include:

- o IP [v4 and v6] label switching
- o L2VPN/PW signaling

To disable unnecessary state exchange for such LDP applications, two new capabilities are being introduced in this document. These new capabilities allow an LDP speaker to notify its LDP peer at the session establishment time when one or more LDP "Non-negotiated applications" are not required/configured on the sender side. Upon receipt of such capability TLV, the receiving LDP speaker MUST disable the advertisement of application state towards the sender. These capabilities can also be sent later in a Capability message to either disable these applications, or to enable previously disabled applications.

#### 3.1. Application Control Capabilities

To control advertisement of state related to non-negotiated LDP applications, namely IP Label switching and L2VPN/PW signaling, two new capability TLVs are defined as described in the following subsections.

##### 3.1.1. IP Label Switching Capability TLV

The IP Label Switching capability is a new Capability Parameter defined with the following format:

0										1										2										3									
0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9
1 0  IP Label Sw. Cap (IANA)																				Length (2)																			
1  Reserved										AF Bitmap																													

The value of the U-bit for the IP capability parameter TLV MUST be set to 1 so that a receiver MUST silently ignore this TLV if unknown

to it, and continue processing the rest of the message. Once advertised, this capability cannot be withdrawn and hence the S-bit must always be set to 1 both in Initialization message and Capability message. The capability data associated with this TLV is 1 byte long "Address Family Bitmap", and hence the TLV length MUST be set to 2.

The Capability data "Address Family Bitmap" is defined as:

```

  7 6 5 4 3 2 1 0
+---+---+---+---+
|   AF bitmap   |
+---+---+---+---+

```

Where:

bit0: IPv4 label switching application

bit1: IPv6 label switching application

bit2-7: Reserved.

A bit in the bitmap is set to 0 or 1 to disable or enable respectively a corresponding IP application.

As described earlier, "IP Label Switching" Capability Parameter TLV MAY be included by an LDP speaker in an Initialization message to signal to its peer LSR that state exchange for IPv4 and/or IPv6 application(s) need to be disabled on a given peer session. This TLV can also be sent later in a Capability message to selectively enable or disable IPv4/v6 label switching application(s).

### 3.1.2. PW Signaling Capability TLV

The "PW Signaling" capability is a new Capability Parameter defined with the following format:

```

      0                               1                               2                               3
0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
|1|0|  PW Sig. Cap (IANA)           |           Length (2)           |
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
|1| Reserved           |E| Reserved           |
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+

```

The value of the U-bit for the PW capability parameter TLV MUST be set to 1 so that a receiver MUST silently ignore this TLV if unknown to it, and continue processing the rest of the message. Once advertised, this capability cannot be withdrawn and hence the S-bit must always be set to 1 in Initialization message or Capability message. The capability data associated with this TLV is 1 byte long and hence the TLV length MUST be set to 2.

The capability data is defined as following byte:

```

7 6 5 4 3 2 1 0
+---+---+---+---+
|E|   Reserved   |
+---+---+---+---+
```

Where E-bit (Enable bit) is used to control PW signaling application by setting it to 0 and 1 to disable and enable the application respectively.

As described earlier, PW Signaling Capability Parameter TLV MAY be included by an LDP speaker in an Initialization message to signal to its peer LSR that state exchange for PW application need to be disabled on given peer session. This TLV can also be sent later in a Capability message to selectively enable/disable the PW Signaling application.

### 3.2. Procedures for Application Control Capabilities in an Initialization message

LDP Capabilities [RFC5561] dictate that the S-bit of capability parameter in an Initialization message MUST be set to 1 and SHOULD be ignored on receipt.

An LDP speaker determines (e.g. via some local configuration or default policy) if they need to disable IP and/or L2VPN/PW applications with a peer LSR. If there is a need to disable, then the IP and/or PW application capability TLVs need to be included in the Initialization message with respective application bits set to 0 to indicate application disable, where the application bit refers to a bit in "Address Family Bitmap" of the "IP Label Switching" Capability or E-bit in "PW Signaling" Capability.

An LDP speaker that supports the "IP Label Switching" and/or "PW Signaling" capability MUST interpret those TLVs in a received Initialization message such that it disables the advertisement of the

application state towards the sender LSR for IP (v4 and/or v6) and/or L2VPN/PW applications if their application control bits are set to 0. If a receiving LDP speaker does not understand the capability TLVs, then it MUST respond to the sender with "Unsupported TLV" Notification as described in LDP Capabilities [RFC5561]. Upon receipt of such Notification, the sender MAY still continue to block/disable its outbound state advertisement towards the peer for the requested disabled applications.

Once this capability has been sent by sender LSR and received and understood by the receiver LSR, then both these LSRs MUST NOT exchange any state related to the disabled applications until and unless these applications are explicitly enabled again (e.g. via the same Capability TLV sent in a Capability message with corresponding application control bit set to 1).

"IP Label Switching" and "PW Signaling" capability TLVs are unilateral/uni-directional in nature. This means that the receiving LSR may not need to send a similar capability TLV in an Initialization or Capability message towards the sender. This unilateral behavior also conforms to the procedures defined in the Section 6 of LDP Capabilities [RFC 5561].

### 3.3. Procedures for Application Control capabilities in a Capability message

If the LDP peer supports "Dynamic Announcement Capability" [RFC5561], then an LDP speaker can send IP Label Switching and/or PW Signaling capability in a Capability message. Once advertised, these capabilities cannot be withdrawn and hence the S-bit of the TLV MUST be set to 1 when sent in a Capability message.

An LDP speaker may decide to send this TLV towards an LDP peer if any of its IP and/or L2VPN/PW signaling applications gets disabled or if previously disabled IP or L2VPN/PW application(s) gets enabled again. In this case, LDP speaker constructs the TLVs with appropriate application control bitmap and sends the corresponding capability TLVs in a Capability message. Furthermore, the LDP speaker also withdraws application(s) related advertised state (such as label bindings) from its peer.

Upon receipt of those TLVs in a Capability message, the receiving LDP speaker reacts in the same manner as it reacts upon the receipt of those TLVs in an Initialization message. Additionally, the receiving LDP speaker withdraws the application(s) related advertised state (such as label bindings) from the sending LDP speaker. If the receiving LDP speaker does not understand or support either Dynamic

Announcement capability or received Application Control capability TLV ("IP Label Switching" or "PW Signaling"), it MUST respond with "Unsupported Capability" notification to the sender of the Capability message.

#### 4. Operational Examples

##### 4.1. Disabling IP/PW label applications on an ICCP session

Consider two PE routers, LSR1 and LSR2, which understand/support "IP Label Switching" and "PW Signaling" capability TLVs. These LSR have an established LDP session due to ICCP application in order to exchange ICCP state related to dual-homed devices connected to these LSRs. Let us assume that LSR1 is provisioned not to exchange any label bindings related to IP (v4/v6) prefixes and PW layer2 FEC (FEC128/129) with LSR2.

To indicate its "disability" for the IP/PW applications, the LSR1 will include both the "IP Label Switching" capability TLV (with bit0-1 of "Address Family Bitmap" set to 0) and "PW Signaling" capability TLV (with E-bit set to 0) in the Initialization message. Upon receipt of those TLVs in Initialization message, the LSR2 will disable any IP/PW address/label binding state advertisement towards LSR1.

The LSR1 will also disable any IP/PW address/label binding state towards LSR2, irrespective of the fact whether or not LSR2 could disable the corresponding application state advertisement towards LSR1.

##### 4.2. Disabling IP Label Switching application on a L2VPN/PW session

Now, consider LSR1 and LSR2 have an established session due to L2VPN/PW application in order to exchange PW (FEC128/129) label bindings for VPWS/VPLS services amongst them. Since in most typical deployments, there is no need to exchange IP (v4/v6) address/label bindings amongst the PE LSRs, let us assume that LSR1 is provisioned to disable IP (v4/v6) application on given PW session towards LSR2.

To indicate its disability for IP application, the LSR1 will include the "IP Label Switching" capability TLV in the Initialization message with bit0-1 (IPv4, IPv6) in "Address Family Bitmap" set to zero. Upon receipt of this TLV in Initialization message, the LSR2 will disable any IP address/label binding state advertisement towards LSR1.

The LSR1 will also disable any IP address/label binding state towards LSR2, irrespective of the fact whether or not LSR2 could disable the corresponding IP application state advertisement towards LSR1.

#### 4.3. Disabling IP application dynamically on an established IP/PW session

Assume that LSRs from previous sections were initially provisioned to exchange both IP and PW state over the session between them, and also support "Dynamic Announcement" capability [RFC5561]. Now, assume that LSR1 is provisioned to disable IP label switching application with LSR2. In this case, LSR1 will first withdraw all its IP label state by sending a single Label Withdraw message with IP prefix Typed Wildcard FEC using the mechanics described in [RFC5918], and Address Withdraw message to withdraw its addresses. LSR1 will also send IP Label Switching capability TLV in Capability message towards LSR2 with bit0-1 (IPv4, IPv6) in "Address Family Bitmap" set to zero. Upon receipt of this TLV, LSR2 will also disable IP application towards LSR1 and withdraw all previous IP application label/address state using the same mechanics as described earlier for LSR1. The disability of IP label switching dynamically should not impact L2VPN/PW application on given session, and both LSRs should continue to exchange PW Signaling application related state.

#### 5. Security Considerations

The proposal introduced in this document does not introduce any new security considerations beyond that already apply to the base LDP specification [RFC5036] and [RFC5920].

#### 6. IANA Considerations

The document introduces following two new capability parameter TLVs and requests following LDP TLV code point assignment by IANA:

- o "IP Label Switching" Capability TLV (requested codepoint: 0x50C)
- o "PW Signaling" Capability TLV (requested codepoint: 0x50D)

## 7. Conclusions

The document proposed a solution using LDP Capabilities [RFC5561] mechanics to disable unnecessary state exchange, if/as desired, between LDP peers for currently non-negotiated IP/PW applications.

## 8. References

### 8.1. Normative References

- [RFC5561] Thomas, B., Raza, K., Aggarwal, S., Aggarwal, R., and Le Roux, J.L., "LDP Capabilities", RFC 5561, July 2009.
- [RFC5918] Asati, R., Minei, I., and Thomas, B. "Label Distribution Protocol Typed Wildcard FEC", RFC 5918, August 2010.
- [ICCP] Martini, L., Salam, S., and Matsushima, S., "Inter-Chassis Communication Protocol for L2VPN PE Redundancy", draft-ietf-pwe3-iccp-03.txt, Work in Progress, July 2010.
- [MLDP] Minei, I., Kompella, K., Wijnands, I., and Thomas, B., "LDP Extensions for Point-to-Multipoint and Multipoint-to-Multipoint Label Switched Paths", draft-ietf-mpls-ldp-p2mp-10.txt, Work in Progress, July 2010.
- [RFC4447] L. Martini, Editor, E. Rosen, El-Aawar, T. Smith, G. Heron, "Pseudowire Setup and Maintenance using the Label Distribution Protocol", RFC 4447, April 2006.
- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC2119, March 1997.

### 8.2. Informative References

- [RFC5036] Andersson, L., Menei, I., and Thomas, B., Editors, "LDP Specification", RFC 5036, September 2007.
- [RFC5920] Fang, L. et al., "Security Framework for MPLS and GMPLS Networks", RFC 5920, July 2010.

## 9. Acknowledgments

The authors would like to thank Eric Rosen for his valuable input and comments.

This document was prepared using 2-Word-v2.0.template.dot.

Authors' Addresses

Kamran Raza  
Cisco Systems, Inc.,  
2000 Innovation Drive,  
Kanata, ON K2K-3E8, Canada.  
E-mail: skraza@cisco.com

Sami Boutros  
Cisco Systems, Inc.  
3750 Cisco Way,  
San Jose, CA 95134, USA.  
E-mail: sboutros@cisco.com

MPLS Working Group  
Internet-Draft  
Intended status: Standards Track  
Expires: April 18, 2011

M. Xiao, Ed.  
L. Jin  
B. Wu  
J. Yang  
ZTE Corporation  
October 15, 2010

Throughput Estimation for MPLS based Transport Networks  
draft-xiao-mpls-tp-throughput-estimation-01

Abstract

An important Operation, Administration and Maintenance requirement of the MPLS Transport Profile (MPLS-TP) is the ability to estimate the throughput (i.e. bandwidth) for an MPLS-TP connection which could be an MPLS-TP PW, LSP or Section. This document specifies OAM packets and protocol mechanisms to facilitate the efficient and precise measurement of throughput.

Status of this Memo

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

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

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

This Internet-Draft will expire on April 18, 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.

## Table of Contents

1. Introduction . . . . .	3
1.1. Conventions . . . . .	3
1.2. Abbreviations . . . . .	3
2. Overview . . . . .	4
2.1. Two-way Throughput Measurement . . . . .	4
2.2. One-way Throughput Measurement . . . . .	5
2.3. Unidirectional Connections . . . . .	5
3. Packet Format . . . . .	5
3.1. Throughput Measurement Indication Packet Format . . . . .	5
3.2. Throughput Measurement Test Packet Format . . . . .	10
4. Throughput Measurement Procedures . . . . .	10
4.1. Transmitting a Throughput Measurement Start Request . . . . .	10
4.2. Receiving a Throughput Measurement Start Request . . . . .	11
4.3. Transmitting a Throughput Measurement Start Reply . . . . .	11
4.4. Receiving a Throughput Measurement Start Reply . . . . .	11
4.5. Sending and Receiving Test Traffic . . . . .	12
4.6. Transmitting a Throughput Measurement Stop Request . . . . .	12
4.7. Receiving a Throughput Measurement Stop Request . . . . .	12
4.8. Transmitting a Throughput Measurement Stop Reply . . . . .	13
4.9. Receiving a Throughput Measurement Stop Reply . . . . .	13
4.10. Consequent Actions and Searching Algorithm . . . . .	13
5. Throughput Measurement Time . . . . .	15
6. Open Issue . . . . .	15
7. IANA Considerations . . . . .	15
8. Security Considerations . . . . .	15
9. Acknowledgements . . . . .	16
10. References . . . . .	16
10.1. Normative References . . . . .	16
10.2. Informative References . . . . .	16
Authors' Addresses . . . . .	17

## 1. Introduction

As defined in [RFC5860], the MPLS-TP OAM toolset MUST provide a function to enable conducting diagnostic tests on a PW, LSP or Section, this function SHOULD be performed on-demand and one example of such diagnostic test consists in estimating the bandwidth of e.g., an LSP.

To make this requirement clearer and provide more details, this sub-function of diagnostic tests is specified as "throughput estimation" in [I-D.ietf-mpls-tp-oam-framework], throughput estimation is an on-demand out-of-service function, that allows verifying the bandwidth/throughput of an MPLS-TP transport path (LSP or PW) before it is put in-service. Throughput estimation is performed between MEPs and can be performed in one-way or two-way mode.

This document specifies the OAM packets and procedures for both one-way and two-way throughput estimation/measurement.

### 1.1. Conventions

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].

### 1.2. Abbreviations

CRC: Cyclic Redundancy Check

G-ACh: Generic Associated Channel

DUT: Device Under Test

LSP: Label Switched Path

MEG: Maintenance Entity Group

MEP: Maintenance Entity Group End Point

MPLS-TP: MPLS Transport Profile

NMS: Network Management System

OAM: Operations, Administration and Maintenance

PHB: Per-hop Behavior

PRBS: Pseudo-Random Bit Sequence

PW: PseudoWire

TLV: Type Length Value

## 2. Overview

In [RFC1242], the throughput is specified as a performance metric for network interconnection device, and it's defined by "the maximum rate at which none of the offered frames are dropped by the device". In MPLS-TP context the concept of throughput is not just for a particular device, but extended to apply to an MPLS-TP connection which could be a PW, LSP or Section.

In [RFC2544], corresponding to [RFC1242], the throughput measurement procedures are specified as "send a specific number of frames at a specific rate through the DUT and then count the frames that are transmitted by the DUT. If the count of offered frames is equal to the count of received frames, the fewer frames are received than were transmitted, the rate of the offered stream is reduced and the test is rerun. The throughput is the fastest rate at which the count of test frames transmitted by the DUT is equal to the number of test frames sent to it by the test equipment". But in current practical throughput measurement scenario, usually the throughput is measured by test equipment using the more efficient and precise binary search algorithm.

It should also be noted that for different test packet size, or test packet pattern, or test packet PHB, or expected measurement resolution, or even sending duration of test traffic, different result of throughput measurement may be obtained, so all these parameters need to be configurable for throughput measurement.

### 2.1. Two-way Throughput Measurement

For a bidirectional MPLS-TP connection, two-way throughput measurement needs to be supported. Two-way throughput should include both the throughput for the forward direction of the connection and the throughput for the reverse direction of the connection. In order to simplify the implementation and facilitate the results collection, all computational overhead and procedures control will be taken by the initiator MEP of throughput measurement, and the peer MEP will act just as a responder. Also note that both the initiator MEP and the peer MEP need to send test traffic for two-way throughput measurement.

It is worth noting that there is another optional definition of two-way throughput estimation, in which only the initiator MEP needs to

send test traffic and the peer MEP will loop back all received test packets. But note that in this case only the minimum of available throughput of the two directions can be achieved, so this optional definition of two-way throughput estimation is not recommended in this draft.

## 2.2. One-way Throughput Measurement

For a bidirectional MPLS-TP connection, one-way throughput measurement also needs to be supported. One-way throughput only indicates the throughput for the forward direction of the connection. Similar to two-way throughput measurement, the initiator MEP controls the whole process of one-way throughput measurement and the peer MEP will act just as a responder. Also note that only the initiator MEP needs to send test traffic for one-way throughput measurement.

## 2.3. Unidirectional Connections

For a unidirectional MPLS-TP connection (such as a unidirectional LSP), only one-way throughput measurement needs to be supported. If it's a unidirectional connection with return path, the procedures of one-way throughput measurement for bidirectional connection still apply. Else if it's a unidirectional connection without return path, the procedures of one-way throughput measurement are not as automatic as that for bidirectional connection, and manual provision of test parameters is needed for every run of sending test traffic. Besides, in this case the peer MEP instead of the initiator MEP will act as calculator for the packet loss of every run.

## 3. Packet Format

For throughput measurement the specific packets sent by the MEP can be divided into indication packets and test packets. The throughput measurement indication packets flow over the Generic Associated Channel Channel (G-ACh) [RFC5586] of an MPLS-TP connection and perform signaling between the initiator MEP and the peer MEP, while the throughput measurement test packets compose the test traffic which intends to emulate the real user traffic.

### 3.1. Throughput Measurement Indication Packet Format

The format of a throughput measurement indication packet is shown below.

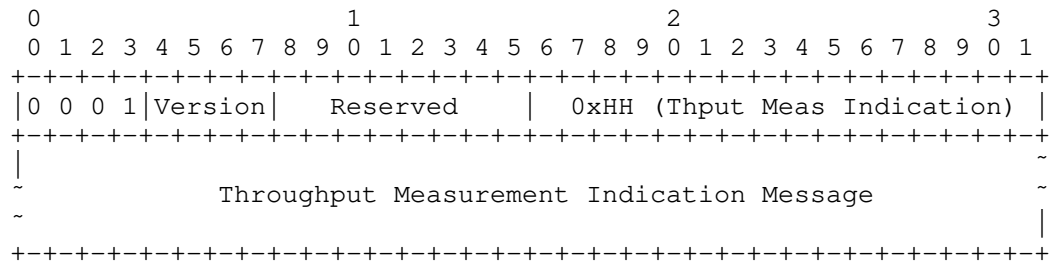


Figure 1: Throughput Measurement Indication Packet

The Version and Reserved field are always set to 0.

The Thput Meas Indication Channel Type is 0xHH (to be assigned by IANA).

The format of a throughput measurement indication message is shown below.

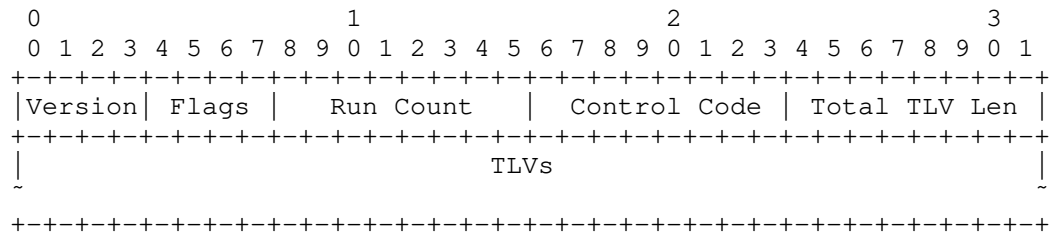


Figure 2: Throughput Measurement Indication Message

Version

The Version Number is currently set to 0.

Flags

Each bit indicates a message control flag. Three flags are defined and listed from left to right as follow:

W	S	R	E
---	---	---	---

W-flag: This Flag represents the operational mode which could be One-way mode or Two-way mode. Set to 0 for a One-way throughput measurement; Set to 1 for a Two-way throughput measurement.

S-flag: This Flag represents the message type which could be Start type or Stop type. Set to 0 for a Start message; Set to 1 for a Stop message.

R-flag: This Flag represents the message direction which could be Forward direction (i.e. Request) or Reverse direction (i.e. Reply). Set to 0 for a Request message; Set to 1 for a Reply message.

E bit (the fourth bit): Reserved for future use and set to 0.

#### Run Count

The Run Count is set to the number of all run times in one throughput measurement process and it starts from 1.

#### Control Code

According to the value of R-flag, the Control Code is set as follow.

For a Request:

0x0: Request (in-band reply requested). Indicates that this request has been sent over a bidirectional connection and the reply is expected over the same connection.

0x1: Request (out-of-band reply requested). Indicates that the reply is expected over an out-of-band path.

0x2: Request (no reply requested). Indicates that no reply is expected.

For a Reply:

0x0: Success. Indicates that the operation succeeded.

0x1: Error. Indicates that the operation failed.

#### Total TLV Length

The total TLV length is the total of all included TLVs.

#### TLVs

According to the values of W-flag, S-flag and R-flag, the TLVs are defined as follow.

For Start Request/Reply message in One-way throughput measurement:

No TLVs are defined at this time.

For Start Request/Reply message in Two-way throughput measurement:

One TLV is defined as follow.

0										1										2										3									
0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9	0	1								
Type = 0										Length = 10																													
Sending Rate																																							
Sending Duration										Packet Size																													
Packet Pattern										PHB										Reserved																			

All the values in this TLV are test parameters for the peer MEP to send test traffic.

#### Sending Rate

The Sending Rate in Mbps is set to the provisioned initial sending rate of test traffic for the first run, and set to the calculated sending rate of test traffic for the rerun.

#### Sending Duration

The Sending Duration in seconds is set to the provisioned sending interval of test traffic for every run.

### Packet Size

The Packet Size in octets is set to the provisioned throughput measurement test packet size.

### Packet Pattern

The Packet Pattern is set to the provisioned throughput measurement test packet pattern. According to [ITU-T Y.1731], four pattern types of throughput measurement test packets pattern types are defined as below:

0x00: Null (all-zeros) signal without CRC-32

0x01: Null (all-zeros) signal with CRC-32

0x02: PRBS ( $2^{31}-1$ ) without CRC-32

0x03: PRBS ( $2^{31}-1$ ) with CRC-32

0x04~0xFF: Reserved for future standardization

### PHB

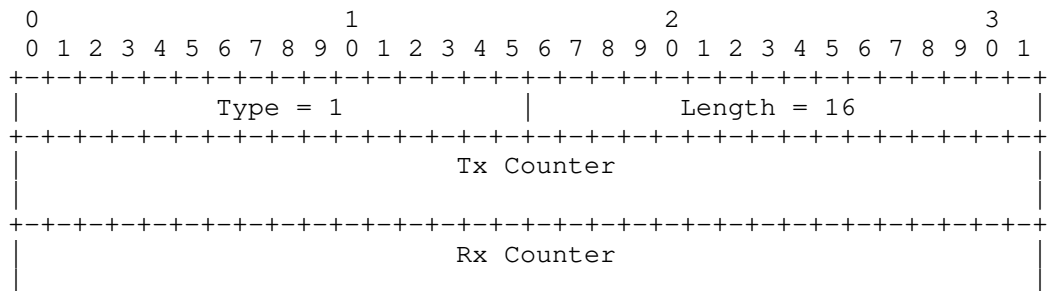
The PHB is set to the provisioned throughput measurement test packet PHB.

### Reserved

Reserved bits for future use and always set to 0.

For Stop Request/Reply message in One-way/Two-way throughput measurement:

One TLV is defined as follow.



[illegible]

## Tx Counter

Tx Counter is set to the number of throughput measurement test packets sent by the local MEP (i.e. the MEP sending this indication message) in this run.

## Rx Counter

Rx Counter is set to the number of throughput measurement test packets received by the local MEP (i.e. the MEP sending this indication message) in this run.

### 3.2. Throughput Measurement Test Packet Format

In order to simplify the implementation of MPLS-TP OAM functions, the format of a throughput measurement test packet should be aligned with the format of a data plane loopback test packet, which is specified in section 5 of [I-D.ietf-mpls-tp-li-lb]. As indicated in section 3.1, four pattern types of throughput measurement test packets can be constructed based on Padding field and optional CRC-32 field.

#### 4. Throughput Measurement Procedures

As specified in [I-D.ietf-mppls-tp-oam-framework], before throughput measurement is initiated, the diagnosed MEG should be put into a Lock status, and an MEG can be put into a Lock status either via NMS action or using the Lock Instruct OAM tool which is specified in [I-D.ietf-mppls-tp-li-lb]. In addition, the test parameters for sending test traffic need to be provisioned at the initiator MEP before initiating a throughput measurement, and they include initial sending rate, sending duration for every run, test packet size, test packet Pattern, test packet PHB and also the expected measurement resolution. Also note that no any provision is needed at the peer MEP.

#### 4.1. Transmitting a Throughput Measurement Start Request

After initiating a throughput measurement operation, the initiator MEP will at first transmit a throughput measurement Start Request to the peer MEP. Also note that for every rerun of sending test traffic, the initiator MEP must transmit this message as beginning.

For one-way throughput measurement, this message is intended to inform the peer MEP about the start of test traffic sending and

trigger the peer MEP to start counting test packets. Specifically if one-way throughput measurement is performed on a unidirectional MPLS-TP connection without return path, the initiator MEP also should start sending test traffic a while (such as 1 second) after transmitting the Start Request. For two-way throughput measurement, except for the same intention as one-way throughput measurement, this message is also intended to convey necessary test parameters to the peer MEP and trigger the test traffic sending at the peer MEP, and note that for rerun only Run Count and Sending Rate in the message need to be changed while other parameters retain initial values. Furthermore, for both one-way and two-way measurement, the initiator MEP should start counting test packets as soon as it transmits this message.

Specifically if the connection is unidirectional, then the Control Code in the message must not be set to 0x0 (in-band reply requested), moreover if no return path exists the Control Code in the message must be set to 0x2 (no reply requested).

#### 4.2. Receiving a Throughput Measurement Start Request

Upon the reception of a throughput measurement Start Request, the peer MEP must inspect this message at first, if no unexpected field or value is found then the peer MEP should start counting test packets. In addition, if the received W-flag indicates that this is a two-way throughput measurement, then the peer MEP also should start sending test traffic.

Specifically if the received W-flag indicates that this is a one-way throughput measurement, and the received Control Code is set to 0x2 (no reply requested) which means the connection is unidirectional without return path, then the peer MEP won't transmit Start Reply.

#### 4.3. Transmitting a Throughput Measurement Start Reply

When the Control Code in a received Start Request is set to 0x0 (in-band reply requested) or 0x1 (out-of-band reply requested), the peer MEP must transmit a throughput measurement Start Reply to the initiator MEP. The Control Code in Start Reply Message should be set to 0x0 to reflect the successful operation at the peer MEP, or on the contrary set to 0x1 to reflect the failed operation at the peer MEP. Except the R-flag and Control Code field, other fields of Start Reply Message will be copied from the received Start Request Message.

#### 4.4. Receiving a Throughput Measurement Start Reply

Upon the reception of a throughput measurement Start Reply, the initiator MEP must inspect this message at first, if no unexpected

field or value is found, and the received Control Code indicates successful operation at the peer MEP, then the initiator MEP should start sending test traffic. If there is no any throughput measurement Start Reply received after a while (such as 1 second), then specific error should be returned at the initiator MEP and no test traffic will be sent from the initiator MEP.

#### 4.5. Sending and Receiving Test Traffic

From above procedures it can be seen that for two-way throughput measurement the pair of MEPs will send test traffic asynchronously, and the peer MEP will start/stop sending test traffic some earlier than the initiator MEP, but the asynchronism has no side-effect on the measurement result because both MEPs shall start counting test packets before they receive any test traffic.

Also note that when the initiator MEP sends test traffic the test parameters are all derived from the provisioned test parameters for the first run, and for rerun only the sending rate is changed and derived from the local calculation. When the peer MEP sends test traffic, the test parameters are all derived from the received Start Request Message.

#### 4.6. Transmitting a Throughput Measurement Stop Request

For every run, after the initiator MEP finished sending test traffic, it will transmit a throughput measurement Stop Request to the peer MEP. This message is intended to inform the peer MEP about the stop of test traffic sending, and also trigger the peer MEP to stop counting test packets and feed back the counters.

Specifically if the connection is unidirectional, then the Control Code in the message must not be set to 0x0 (in-band reply requested), moreover if no return path exists the Control Code in the message must be set to 0x2 (no reply requested).

#### 4.7. Receiving a Throughput Measurement Stop Request

Upon the reception of a throughput measurement Stop Request, the peer MEP must inspect this message at first, if no unexpected field or value is found then the peer MEP should stop counting test packets.

Specifically if the received W-flag indicates that this is a one-way throughput measurement, and the received Control Code is set to 0x2 (no reply requested) which means the connection is unidirectional without return path, then the peer MEP won't transmit Stop Reply and it will use the received Tx Counter to calculate test packet loss directly.

#### 4.8. Transmitting a Throughput Measurement Stop Reply

When the Control Code in a received Stop Request is set to 0x0 (in-band reply requested) or 0x1 (out-of-band reply requested), the peer MEP must transmit a throughput measurement Stop Reply to the initiator MEP. The Control Code in Stop Reply Message should be set to 0x0 to reflect the successful operation at the peer MEP, or on the contrary set to 0x1 to reflect the failed operation at the peer MEP. Furthermore, the Stop Reply is transmitted also to confirm that the peer MEP has stopped sending test traffic for this run. The Tx Counter and Rx Counter are set to the test packet counting values at the peer MEP.

#### 4.9. Receiving a Throughput Measurement Stop Reply

Upon the reception of a throughput measurement Stop Reply, the initiator MEP must inspect this message at first, if no unexpected field or value is found, and the received Control Code indicates successful operation at the peer MEP, then the initiator MEP should stop counting test packets and start calculating the test packet loss. Suppose the Tx Counter and Rx Counter for the initiator MEP are TxP1 and RxP1, and for the peer MEP are TxP2 and RxP2.

For two-way throughput measurement, the calculation formulas are as follow:

$$\text{Packet Loss (forward)} = \text{TxP1} - \text{RxP2}$$
$$\text{Packet Loss (reverse)} = \text{TxP2} - \text{RxP1}$$

For one-way throughput measurement, the calculation formula is as follow:

$$\text{Packet Loss (one-way)} = \text{TxP1} - \text{RxP2}$$

If there is no any throughput measurement Stop Reply received after a while (such as 1 second), then specific error should be returned at the initiator MEP and no consequent action will happen.

#### 4.10. Consequent Actions and Searching Algorithm

Procedures for one run of test traffic sending and test packet loss calculation have been described above in details, but usually iterative reruns of the procedures are needed for a throughput measurement. Whether the rerun is needed or not is based on the calculated test packet loss and whether the expected measurement resolution is met. For one-way throughput measurement, if calculated Packet Loss (one-way) is equal to zero and the expected measurement

resolution is met, then rerun is not needed (i.e. the one-way throughput measurement finished) and the current sending rate is the measured one-way throughput, otherwise the one-way throughput measurement proceeds. For two-way throughput measurement, if calculated forward Packet Loss and reverse Packet Loss are both equal to zero and the expected measurement resolution for both forward and reverse directions is met, then rerun is not needed (i.e. the two-way throughput measurement finished) and the current sending rate for forward/reverse direction is the measured forward/reverse throughput, otherwise the two-way throughput measurement proceeds, and in this case the sending rates for rerun should be calculated for forward direction and reverse direction respectively.

The simple and efficient binary search algorithm is RECOMMENDED to calculate the sending rate for the next run, which is the only changed test parameter compared with this run. How the binary search works, if packet loss is found for this run, it searches downwards for a lower rate which is halfway rate between the rate of this run and the known highest rate at which no packet loss is found; if no packet loss is found but the expected measurement resolution is not met for this run, it searches upwards for a higher rate which is halfway rate between the rate of this run and the known lowest rate at which packet loss is found; the measurement searches among higher and lower rates on the analogy of this, until it finds the rate at which no test packet is lost and expected measurement resolution is met, and this rate is the measured throughput. How to judge whether the expected measurement resolution is met or not, if the rate difference between the two consecutive runs (i.e. this run and the previous run), expressed as a percentage, is smaller than or equal to the specified measurement resolution, it's known as that the expected measurement resolution is met, otherwise it's not met.

For example, suppose to measure the throughput of a connection whose actual throughput is 70Mbps, the provisioned initial sending rate is 100Mbps and the specified measurement resolution is 0.1. Note that the initial sending rate should be higher than the actual throughput, otherwise the binary search is not applicable, and so it's often set to the maximum theoretical throughput of the measured connection. For the first run, packet loss is found, so for the second run, the sending rate will be calculated as  $(100+0)/2 = 50\text{Mbps}$ , no packet loss is found, then the resolution will be calculated as  $(100-50)/50 = 1$ , which is bigger than 0.1, the expected measurement resolution is not met, so for the third run, the sending rate will be calculated as  $(100+50)/2 = 75\text{Mbps}$ , packet loss is found, so for the fourth run, the sending rate will be calculated as  $(50+75)/2 = 62.5\text{Mbps}$ , no packet loss is found, then the resolution will be calculated as  $(75-62.5)/62.5 = 0.2$ , which is bigger than 0.1, the expected measurement resolution is not met, so for the fifth run, the sending rate will be

calculated as  $(75+62.5)/2 = 68.75\text{Mbps}$ , no packet loss is found, then the resolution will be calculated as  $(68.75-62.5)/68.75 = 0.09$ , which is smaller than 0.1, the expected measurement resolution is met, so the measurement finished and the rate 68.75Mbps is the measured throughput.

Other algorithms than the binary search algorithm could also be used to search throughput in practice, e.g. increasing or decreasing the sending rate in a fixed step from a specified initial sending rate until the test packet loss appears or disappears.

## 5. Throughput Measurement Time

The throughput measurement time is about the product of sending duration for one run and number of all run times. The sending duration for one run is provisioned before the throughput measurement starts, and the number of all run times is related to several factors, which include the provisioned initial sending rate, the applied searching algorithm and the specified expected measurement resolution. It's obvious that longer sending duration is provisioned, then longer throughput measurement time is needed, but it should be noted that longer sending duration can result in more precise measured throughput, so there should be a balance between them. Also obviously the expectations for shorter throughput measurement time and higher throughput measurement resolution are mutually exclusive, so the balance between them is needed too.

## 6. Open Issue

Wouldn't it be better to have a threshold on the acceptable frame loss rate and not require absolutely no packet loss?

[Editor's note: As the authors know in practice when the throughput is measured by test devices, one threshold on the acceptable frame loss rate is configurable, but in [RFC1242] and [RFC2544] the throughput is defined as that way no packet loss permitted.]

## 7. IANA Considerations

To be added in a later version of this document.

## 8. Security Considerations

To be added in a later version of this document.

## 9. Acknowledgements

The authors would like to thank Huub (Huawei), Curtis (Infinera) and Ayal (celtro) for their valuable comments on this draft.

## 10. References

### 10.1. Normative References

- [I-D.ietf-mpls-tp-li-lb]  
Boutros, S., Sivabalan, S., Swallow, G., Bryant, S., and C. Pignataro, "MPLS Transport Profile Lock Instruct and Loopback Functions", draft-ietf-mpls-tp-li-lb-00 (work in progress), September 2010.
- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC5586] Bocci, M., Vigoureux, M., and S. Bryant, "MPLS Generic Associated Channel", RFC 5586, June 2009.
- [RFC5860] Vigoureux, M., Ward, D., and M. Betts, "Requirements for Operations, Administration, and Maintenance (OAM) in MPLS Transport Networks", RFC 5860, May 2010.

### 10.2. Informative References

- [I-D.ietf-mpls-tp-oam-framework]  
Allan, D., Busi, I., Niven-Jenkins, B., Fulignoli, A., Hernandez-Valencia, E., Levrau, L., Sestito, V., Sprecher, N., Helvoort, H., Vigoureux, M., Weingarten, Y., and R. Winter, "Operations, Administration and Maintenance Framework for MPLS- based Transport Networks", draft-ietf-mpls-tp-oam-framework-09 (work in progress), October 2010.
- [ITU-T Y.1731]  
International Telecommunications Union - Telecommunication Standardization, "OAM functions and mechanisms for Ethernet based networks", ITU-T Y.1731, February 2008.
- [RFC1242] Bradner, S., "Benchmarking terminology for network interconnection devices", RFC 1242, July 1991.
- [RFC2544] Bradner, S. and J. McQuaid, "Benchmarking Methodology for Network Interconnect Devices", RFC 2544, March 1999.

Authors' Addresses

Min Xiao (editor)  
ZTE Corporation

Email: xiao.min2@zte.com.cn

LiZhong Jin  
ZTE Corporation

Email: lizhong.jin@zte.com.cn

Bo Wu  
ZTE Corporation

Email: wu.bo@zte.com.cn

Jian Yang  
ZTE Corporation

Email: yang\_jian@zte.com.cn



MPLS Working Group  
Internet-Draft  
Intended status: Standards Track  
Expires: April 28, 2011

F. Zhang, Ed.  
L. Jin  
B. Wu  
ZTE Corporation  
October 25, 2010

The Analysis of MPLS-TP Path Segment Monitoring  
draft-zhang-mpls-tp-path-segment-monitoring-01

Abstract

This specification analyzes the different schemes to realize path segment monitoring in MPLS-TP network.

Status of this Memo

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

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

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

This Internet-Draft will expire on April 28, 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.

## Table of Contents

1. Introduction . . . . .	3
2. Conventions used in this document . . . . .	3
3. Path Segment Monitoring Analysis . . . . .	3
3.1. MBB . . . . .	3
3.2. Local Rerouting . . . . .	4
3.3. TTL TLV . . . . .	5
3.3.1. The scaling Analysis . . . . .	6
4. IANA Considerations . . . . .	6
5. Security Considerations . . . . .	6
6. Acknowledgement . . . . .	6
7. Normative references . . . . .	7
Authors' Addresses . . . . .	7

## 1. Introduction

In order to monitor, protect and manage a portion (i.e. segment or concatenated segment) of a transport path, a path segment is defined between the edges of the portion of the LSP that needs to be monitored, protected or managed. If this path segment is created as a hierarchical LSP, it is called SPME (Sub-Path Maintenance Element).

SPMEs are usually instantiated when the transport path is created by either the management plane or control plane for proactive monitoring. However, pre-design and pre-configuration of all the considered patterns of SPME are not sometimes preferable in real operation due to the burden of design works, a number of header consumptions, bandwidth consumption and so on, as described in section 3.8 of [I-D.ietf-mpls-tp-oam-framework].

There are different schemes to configure SPMEs after the transport path has been created, and two network objectives SHOULD be met:

1. The monitoring and maintenance of existing transport paths has to be conducted in service without traffic disruption.
2. The monitored or managed transport path condition has to be exactly the same irrespective of any configurations necessary for maintenance.

Here we will discuss the the advantages and disadvantages of different potential schemes.

## 2. Conventions used in 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 RFC2119.

## 3. Path Segment Monitoring Analysis

### 3.1. MBB

The make-before-break (MBB) procedures which are supported by MPLS allow the creation of a SPME on existing LSPs in-service without traffic disruption, as described in [RFC5921]. An SPME can be defined corresponding to one or more end-to-end LSPs at first, then new end-to-end LSPs that are tunneled within the SPME can be set up, which may require coordination across administrative boundaries, finally traffic of the existing LSPs is switched over to the new end-

to-end tunneled LSPs. The old end-to-end LSPs can then be torn down. See the figure below, copied from [RFC5921]:

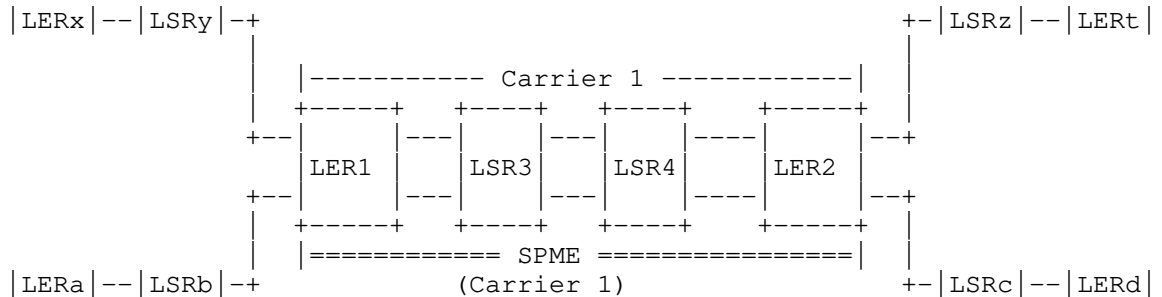


Figure 1: SPME for a set of transport path segments

In the MBB schemes, LER1 needs to inform the old LSPs's ingress nodes (for example, LERx and LERa) that a SPME has been setup to monitor the segment between LER1 and LER2, so that LERx/LERa can instantiate the new LSPs. However, the coordination schemes across administrative boundaries are not explicitly described in [RFC5921].

[RFC4736] gives the RSVP-TE extension to realize reoptimization of MPLS TE Loosely Routed LSP. It is said that when a mid-point LSR whose next hop is a loose hop or an abstract node can locally trigger a path re-evaluation when a configurable timer expires, some specific events occur (e.g., link-up event), or the user explicitly requests it. If a preferable path is found, the LSR sends an RSVP PathErr to the head-end LSR (Error code 25 (Notify), Error sub-code=6 ("preferable path exists")). Although SPME can be seen as a new link, the ingress nodes do know that they need to be triggered to establish new LSPs. In order to differentiate the cases between SPME and reoptimization, the new value "SPME up" is suggested to be assigned.

As we can see, network objective (1) can be met, but network objective (2) can not be met due to the new assignment of MPLS labels.

### 3.2. Local Rerouting

A bidirectional LSP1(LERx-LSRy-LER1-LSR3-LSR4-LER2-LSRz-LERt) exists between LERx and LERt, the forwarding label values along LERx->LERt direction are L<sub>yx</sub>-L<sub>ly</sub>-L<sub>31</sub>-L<sub>43</sub>-L<sub>24</sub>-L<sub>z2</sub>-L<sub>tz</sub>, and the forwarding label values along LERt->LERx direction are L<sub>xy</sub>-L<sub>y1</sub>-L<sub>13</sub>-L<sub>34</sub>-L<sub>42</sub>-L<sub>2z</sub>-L<sub>zt</sub>. Assuming that SPME1 (LER1-LSR3-LSR4- LER2) is established to monitor this LSP, in order to restrict the operation in the scope of Carrier

1, local rerouting technology described in [RFC4090] can be used here.

LER1 uses the label L24 as the inner label and pushes it into SPME1, LER2 uses the label L13 as the inner label and pushes it into SPME1. But LER1 (LER2) needs to learn L24 (L13), which can be learned by the following procedures:

When SPME1 is up, LER1 pushes LSP1's Path message into SPME1, the next hop is changed from LSR3 to LER2, upstream label unchanged (L13 is allocated to LER2). Similarly, LER2 pushes LSP1's Resv message into SPME1, the next hop is changed from LSR4 to LER1, and Label unchanged (L24 is allocated to LER1). After LER1 (LER2) has learned the inner label value L24 (L13), it can push the user traffic into SPME1.

If LSP1 is unidirectional, LER1 pushes LSP1's Path message into SPME1, the next hop is changed from LSR3 to LER2. But for Resv message, the next of LER2 is changed from LSR4 to LER1, and it needs to be transmitted hop by hop.

Local rerouting is more optimized compared to MBB, for the new assigned labels just exist in the scope of Carrier 1, but it still can not fully math the requirements of network objective (2).

### 3.3. TTL TLV

In order to totally meet the requirements of network objective (2), the schemes based on non-label stack are needed.

TTL TLV, as one of the optional ACH TLV objects, is defined in [I-D.boutros-mpls-lsp-ping-ttl-tlv], which is used to inform the receiver how many hops away the originator is on the path of the MS-PW or Bidirectional LSP. It can be used to realize path segment monitoring also, see the figure below.

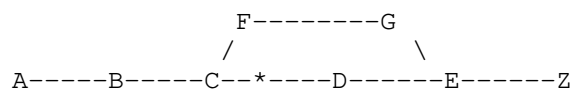


Figure 2: TTL TLV for PSM

The path segment PS1 (C-D-E) of LSP1 (A-B-C-D-E-Z) needs to be provisioned. Node C, as the MEP node of PS1, sends OAM message (like CC/CV, PM loss/dely, etc.) to node D, the TTL TLV MUST be inserted. TTL value is set to the hop counts from E to C, here it is 2 (if LSP1 is an associated bidirectional LSP, the hops form E to C maybe not be

2). In this way, node E can use the hops carried in TTL TLV to response the OAM message.

The TTL values can be configured by NMS, or learned by control plane. [I-D.ietf-mpls-tp-identifiers] describes the MEP-ID of Pseudowire Segments, and the MEP configuration of path Segments can be defined similarly. That is to say, the MEP\_ID of path segment can be formed by a combination of a LSP MEP\_ID and the identification of the local node, such as "Src-Global\_ID::Src-Node\_ID::Src-Tunnel\_Num::LSP\_Num::PS-Global\_ID::PS-Node\_ID".

### 3.3.1. The scaling Analysis

Assuming there is another path segment PS2 that exists between node B and E, proactive and on-demand OAM messages are running between B and E also. Just like PS1, the TTL TLV MUST be inserted, and the value is the hop counts from E to B (here it is 3). At some time, a defect happens between node C and D, the customer traffic would be switched from PS1 to the backup path(C---F---G---E). However, node B may not know that node C has switched all the traffic to the backup path, and in this case, node E can not receive the OAM message sent by node B and may make wrong decision.

In conclusion, TTL TLV scheme can meet both the two network objectives. But it can not be used if two or more path segments are nested.

## 4. IANA Considerations

A new error sub-code values for the RSVP PathErr Notify message (Error code=25) is required in this document:

Error sub-code=TBD by IANA: "SPME up".

## 5. Security Considerations

TBD.

## 6. Acknowledgement

The authors would like to thank Hui Su for the discussion, thank Alexander Vainshtein, Kannan KV Sampath, Nurit Sprecher, Yoshinori Koike for their valuable comments.

## 7. Normative references

- [I-D.boutros-mpls-lsp-ping-ttl-tlv]  
Manral, V., Boutros, S., Sivabalan, S., Saxena, S., and G. Swallow, "Definition of Time-to-Live TLV for LSP-Ping Mechanisms", draft-boutros-mpls-lsp-ping-ttl-tlv-01 (work in progress), June 2010.
- [I-D.ietf-mpls-tp-identifiers]  
Bocci, M. and G. Swallow, "MPLS-TP Identifiers", draft-ietf-mpls-tp-identifiers-02 (work in progress), July 2010.
- [I-D.ietf-mpls-tp-oam-framework]  
Allan, D., Busi, I., Niven-Jenkins, B., Fulignoli, A., Hernandez-Valencia, E., Levrau, L., Sestito, V., Sprecher, N., Helvoort, H., Vigoureux, M., Weingarten, Y., and R. Winter, "Operations, Administration and Maintenance Framework for MPLS- based Transport Networks", draft-ietf-mpls-tp-oam-framework-09 (work in progress), October 2010.
- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC4090] Pan, P., Swallow, G., and A. Atlas, "Fast Reroute Extensions to RSVP-TE for LSP Tunnels", RFC 4090, May 2005.
- [RFC4736] Vasseur, JP., Ikejiri, Y., and R. Zhang, "Reoptimization of Multiprotocol Label Switching (MPLS) Traffic Engineering (TE) Loosely Routed Label Switched Path (LSP)", RFC 4736, November 2006.
- [RFC5921] Bocci, M., Bryant, S., Frost, D., Levrau, L., and L. Berger, "A Framework for MPLS in Transport Networks", RFC 5921, July 2010.

Authors' Addresses

Fei Zhang (editor)  
ZTE Corporation  
4F, RD Building 2, Zijinghua Road  
Yuhuatai District, Nanjing 210012  
P.R.China

Phone: +86 025 52877612  
Email: zhang.fei3@zte.com.cn

LZ Jin  
ZTE Corporation  
889, Bibo Road, Zijinghua Road  
Pudong District, Shanghai 201203  
P.R.China

Phone: +86 021 68896273  
Email: lizhong.jin@zte.com.cn

Bo Wu  
ZTE Corporation  
4F, RD Building 2, Zijinghua Road  
Yuhuatai District, Nanjing 210012  
P.R.China

Phone: +86 025 52877276  
Email: wu.bo@zte.com.cn



Network working group  
Internet Draft  
Intended status: Standards Track  
Updates: RFC 5036 (if approved)  
Expires: April 2011

L. Zheng  
M. Chen  
Huawei Technologies  
October 8, 2010

## LDP Hello Cryptographic Authentication

draft-zheng-mpls-ldp-hello-crypto-auth-00.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 April 8, 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.

## Abstract

This document introduces a new Cryptographic Authentication TLV which is used in LDP Hello message as an optional parameter. It enhances the authentication mechanism for LDP by securing the Hello message against spoofing attack.

## Conventions used in 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 RFC-2119 [RFC2119].

## Table of Contents

1. Introduction.....	2
2. Cryptographic Authentication TLV.....	4
2.1. Optional Parameter for Hello Message.....	4
2.2. Cryptographic Authentication TLV Encoding.....	4
3. Processing Hello Message Using Cryptographic Authentication...	5
3.1. Transmission Using Cryptographic Authentication.....	6
3.2. Receipt Using Cryptographic Authentication.....	6
4. Security Considerations.....	7
5. IANA Considerations.....	7
6. Acknowledgments.....	8
7. References.....	8
7.1. Normative References.....	8
7.2. Informative References.....	8
Authors' Addresses.....	9

## 1. Introduction

The Label Distribution Protocol (LDP) [RFC 5036] utilizes LDP sessions that run between LDP peers. The peers may be directly connected at the link level or may be remote. A label switching router (LSR) that speaks LDP may be configured with the identity of its peers or may discover them using the LDP Hello message sent encapsulated in UDP that may be addressed to "all routers on this subnet" or to a specific IP address. Periodic Hello messages are also used to maintain the relationship between LDP peers necessary to keep the LDP session active.

Unlike all other LDP messages, the Hello messages are sent using UDP not TCP. This means that they cannot benefit from the security mechanisms available with TCP. [RFC5036] does not provide any security mechanisms for use with Hello messages except to note that some configuration may help protect against bogus discovery events.

Spoofing a Hello packet for an existing adjacency can cause the valid adjacency to time out and in turn can result in termination of the associated session. This can occur when the spoofed Hello specifies a smaller Hold Time, causing the receiver to expect Hellos within this smaller interval, while the true neighbor continues sending Hellos at the previously agreed lower frequency. Spoofing a Hello packet can also cause the LDP session to be terminated directly, which can occur when the spoofed Hello specifies a different Transport Address, other than the previously agreed one between neighbors. Spoofed Hello messages is observed and reported as real problem in production networks.

As described in [RFC5036], the threat of spoofed Basic Hellos can be reduced by accepting Basic Hellos only on interfaces to which LSRs that can be trusted, and ignoring Basic Hellos not addressed to the "all routers on this subnet" multicast group. Spoofing attacks via Extended Hellos are potentially more serious threat. An LSR can reduce the threat of spoofed Extended Hellos by filtering them and accepting only those originating at sources permitted by an access list. However, performing the filtering using access lists requires LSR resource, and the LSR is still vulnerable to the IP source address spoofing.

This document introduces a new Cryptographic Authentication TLV which is used in LDP Hello message as an optional parameter. It enhances the authentication mechanism for LDP by securing the Hello message against spoofing attack, and an LSR can be configured to only accept Hello messages from specific peers when authentication is in use.

Using this Cryptographic Authentication TLV, one or more secret keys (with corresponding key IDs) are configured in each system. For each LDP Hello packet, the key is used to generate and verify a "message digest" or "message hash" that is stored in the LDP Hello packet. A sequence number is also carried in each packet to help avoid replay attacks.

## 2. Cryptographic Authentication TLV

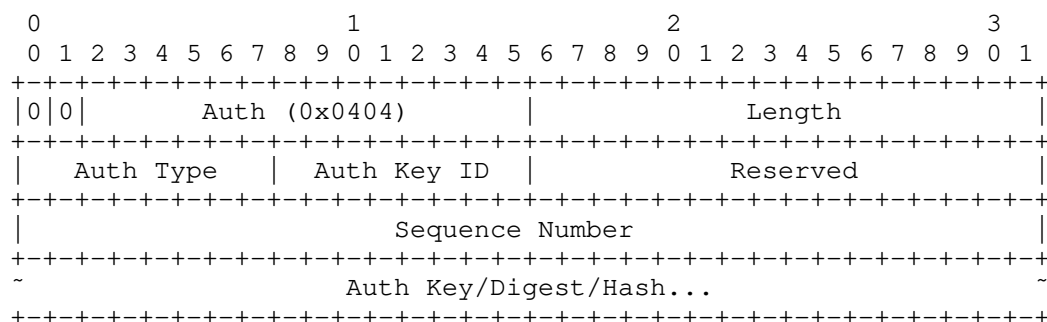
### 2.1. Optional Parameter for Hello Message

[RFC5036] defines the encoding for the Hello message. Each Hello message contains zero or more Optional Parameters, each encoded as a TLV. Three Optional Parameters are defined by [RFC5036]:

Optional Parameter	Type
IPv4 Transport Address	0x0401
Configuration Sequence Number	0x0402
IPv6 Transport Address	0x0403

This document defines a new Optional Parameter: the Cryptographic Authentication parameter. The Cryptographic Authentication TLV Encoding is described in section 2.2.

### 2.2. Cryptographic Authentication TLV Encoding



- Type: 0x0404 (TBD by IANA), Cryptographic Authentication
- Length: Specifying the length in octets of the value field.
- Auth Type: The authentication type in use
  - 0 - Keyed MD5

- 1 - Meticulous Keyed MD5
- 2 - Keyed SHA1
- 3 - Meticulous Keyed SHA1
- 4 - Keyed SHA-256
- 5 - Keyed SHA-384
- 6 - Keyed SHA-512
- 7-255 - Reserved for future use  
(TBD by IANA)

- Auth Key ID: The authentication key ID in use for this packet.  
This allows one or more keys to be active simultaneously.
- Reserved: MUST be set to zero on transmit, and ignored on receipt.
- Sequence Number: The sequence number for this packet, providing protection against replay attacks. The value is incremented occasionally. For Meticulous Keyed MD5 and Meticulous Keyed SHA1 Authentication, this value is incremented for each successive packet transmitted for a session.
- Auth Key/Digest/Hash:

This field carries the MD5/SHA1/SHA2 key digest/hash for the packet. The length of the Auth Key/Digest/Hash varies based on the cryptographic algorithm used, which is shown as below:

Auth type	Length
-----	-----
Keyed MD5	16 bytes
Meticulous Keyed MD5	16 bytes
Keyed SHA1	20 bytes
Meticulous Keyed SHA1	20 bytes
Keyed SHA-256	32 bytes
Keyed SHA-384	48 bytes
Keyed SHA-512	64 bytes

When calculating the digest/hash, the shared key is stored in this field, padding with trailing zeros if needed.

### 3. Processing Hello Message Using Cryptographic Authentication

The Cryptographic Authentication mechanisms described in this draft are very similar to those used in other protocols. One or more secret keys (with corresponding key IDs) are configured in each

system. One of the keys is included in a digest or a hash calculated over the outgoing LDP Hello packet, but the Key itself is not carried in the packet.

A sequence number is also carried in each packet to help avoid replay attacks. The sequence number may be incremented in a circular fashion. For most of the authentication scheme in use, the sequence number is occasionally incremented (The decision as to when to increment the sequence number is implementation dependent and outside the scope of this document). Specifically, for Meticulous Keyed MD5 and Meticulous Keyed SHA1, the sequence number is incremented on every packet.

### 3.1. Transmission Using Cryptographic Authentication

Prior to transmitting Hello message, the Auth Type field is set to indicate the authentication type in use. The Auth Key ID field is set to the ID of the current authentication key. The Sequence Number field is set, possibly having been incremented from the last message sent according to the scheme in place. The authentication key is placed into the Auth Key/Digest field, padding with trailing zeros as necessary, for digest/hash calculation.

An MD5 digest or a SHA1/SHA2 hash is calculated over the entire LDP Hello packet. The resulting digest/hash is stored in the Auth Key/Digest/Hash field prior to transmission. The secret key is replaced by the digest/hash, and MUST NOT be carried in the packet.

### 3.2. Receipt Using Cryptographic Authentication

The receiving LSR applies acceptability criteria for received Hellos using cryptographic authentication. If the Cryptographic Authentication TLV is unknown to the receiving LSR, the received packet MUST be discarded according to Section 3.5.1.2.2 of [RFC5036].

If the Cryptographic Authentication TLV in a received Hello packet does not contain a known and acceptable Auth Type value, then the received packet MUST be discarded. If the Auth Key ID field does not match the ID of a configured authentication key, the received packet MUST be discarded.

For most of the authentication scheme in use, if the received sequence number lies outside of the range of last sequence number received to last sequence number received +(Hello Hold Time/Hello Interval) inclusive, the received packet MUST be discarded. Specifically, for Meticulous Keyed MD5 and Meticulous Keyed SHA1, if the received sequence number lies outside of the range of last sequence number received+1 to last sequence number received +(Hello Hold Time/Hello Interval) inclusive, the received packet MUST be discarded.

The receiving LSR replaces the contents of the Auth Key/Digest/Hash field with the authentication key specified by the received Auth Key ID field. If the MD5 digest or SHA1/SHA2 hash of the entire LDP Hello packet is equal to the received value of the Auth Key/Digest/Hash field, the received packet is accepted for other normal checks and processing as described in [RFC5036]. Otherwise, the received packet MUST be discarded.

#### 4. Security Considerations

Section 1 of this document describes the security issues arising from the use of unsecured LDP Hello messages. In order to combat those issues, it is RECOMMENDED that all deployments use the Cryptographic Authentication TLV to secure the Hello message.

The quality of the security provided by the Cryptographic Authentication TLV depends completely on the strength of the cryptographic algorithm in use, the strength of the key being used, and the correct implementation of the security mechanism in communicating LDP implementations. Also, the level of security provided by the Cryptographic Authentication TLV varies based on the authentication type used.

#### 5. IANA Considerations

IANA maintains a registry of LDP message parameters with a sub-registry to track LDP TLV Types. This document request IANA to assign a new TLV Types as follows:

TLV	Type
Cryptographic Authentication	0x0404 (TBD)

This document also request IANA to assign a new registry titled "LDP Hello Authentication Type", its recommended values as follows:

Value	LDP Hello Authentication Type Name
0	Keyed MD5
1	Meticulous Keyed MD5
2	Keyed SHA1
3	Meticulous Keyed SHA1
4	Keyed SHA-256
5	Keyed SHA-384
6	Keyed SHA-512
7-255	Unassigned

(TBD)

## 6. Acknowledgments

The authors would like to thank Liu Xuehu for his work on background and motivation for LDP Hello authentication. The authors also would like to thank Adrian Farrel, Thomas Nadeau and So Ning for their comments.

## 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.
- [RFC5036] Andersson, L., Minei, I., and B. Thomas, "LDP Specification", RFC 5036, October 2007.

### 7.2. Informative References

- [RFC2385] Heffernan, A., "Protection of BGP Sessions via the TCP MD5 Signature Option", RFC 2385, August 1998.
- [RFC5709] Bhatia, M., Manral, V., Fanto, M., White, R., Barnes, M., Li, T., and R. Atkinson, "OSPFv2 HMAC-SHA Cryptographic Authentication", RFC 5709, October 2009.
- [RFC5880] Katz, D. and D. Ward, "Bidirectional Forwarding Detection", RFC 5880, June 2010.

Authors' Addresses

Lianshu Zheng  
Huawei Technologies Co., Ltd.  
Huawei Building, No.3 Xinxu Road,  
Hai-Dian District,  
Beijing 100085  
China

Email: verozheng@huawei.com

Mach(Guoyi) Chen  
Huawei Technologies Co., Ltd.  
Huawei Building, No.3 Xinxu Road,  
Hai-Dian District,  
Beijing 100085  
China

Email: mach@huawei.com