

Internet Engineering Task Force
Internet-Draft
Intended status: Standards Track
Expires: January 12, 2014

S. Alvarez
K. Raza
S. Boutros
Cisco Systems, Inc.
July 11, 2013

Signaling Color Label Switched Paths Using LDP
draft-alvarez-mpls-ldp-color-lsp-00

Abstract

This document describes extensions to the Label Distribution Protocol (LDP) to signal a switching preference in the presence of multiple paths. A label switched router (LSR) can associate locally a color with one or more downstream paths or links, and signal a label path per color to upstream LSRs. Based on local policy, LSRs can select between these color LSPs to implement a forwarding preference on a downstream LSR. An egress LSR may influence the signaling decision of other LSRs by signaling interest in specific colors.

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 January 12, 2014.

Copyright Notice

Copyright (c) 2013 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	2
1.1. Requirements Language	3
2. Applicability of Colors to FEC Elements	3
3. Establishing Color LSPs	3
4. LDP Color LSP Capability	5
5. Color Lists	6
5.1. Color List TLV	6
5.2. Procedures	7
6. Color Address Families	8
6.1. Color IP Address Family	9
6.2. Color IPv6 Address Family	9
7. Color FECs	10
7.1. Color Prefix FEC Element	10
7.2. Color Multipoint FEC Elements	10
7.3. Procedures	10
8. Other FEC-based Features	10
8.1. Typed Wildcard Forward Equivalence Class	10
8.2. Signaling Convergence (End-of-LIB)	11
8.3. LSP Ping Extensions	11
8.3.1. Color Prefix FEC	11
8.3.2. Color Multipoint FEC	13
9. IANA Considerations	13
10. Security Considerations	14
11. References	14
11.1. Normative References	14
11.2. Informative References	14
Authors' Addresses	15

1. Introduction

The extensions in this document allow LSRs to implement a forwarding preference between different paths available to downstream LSRs. While multiple paths between two LSR may have the same cost from an routing perspective, individual paths may have intrinsic characteristics that an LSR may prefer. As an example, some paths may have a significant difference in terms of latency, reliability, protection, bandwidth or path diversity among other characteristics. An LSR may want to allow upstream LSRs to determine what traffic is forwarded down specific paths.

A sample deployment scenario for color LSPs involves an LDP network using targeted LDP over RSVP-TE LSPs. A given head-end provider (P) device can establish multiple TE LSPs to another tail-end P device. One TE LSP can be engineered for low latency while the other TE LSPs can be engineered for high bandwidth. The head-end P device can associate locally the low-latency TE LSP with one color and the other TE LSPs with a second color. This device can signal two paths (one per color) to upstream LDP peers. With these two paths, a provider edge (PE) device can implement local policies to implement a forwarding preference for low latency or high bandwidth at the downstream P device.

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].

2. Applicability of Colors to FEC Elements

A color LSP can be defined for the following types of LDP FEC elements:

Prefix (0x2)

P2MP (0x6)

MP2MP upstream (0x7)

MP2MP downstream (0x8)

and for the following address families:

IP (version 4) (0x1)

IPv6 (0x2)

3. Establishing Color LSPs

Three LSR roles can be involved in establishing a color LSP:

Color LSR:

A transit node with multiple paths towards a destination. It advertises color paths to other upstream LSRs according to a local forwarding association between colors and downstream paths. Advertisement may occur as response to an egress LSR interest in specific color paths.

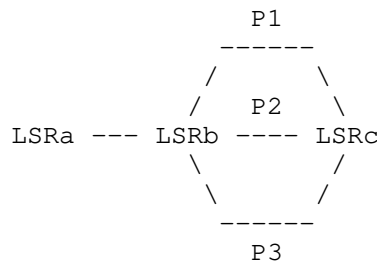
Ingress LSR:

Selects between color paths associated with a given FEC to influence the forwarding by downstream color LSRs.

Egress LSR:

May signal interest in particular colors towards upstream LSRs in order to influence the signaling behavior of color LSRs.

Figure 1 illustrates an example of a network implemententing a forwarding preference policy for a prefix FEC. In this example, LSRa is an ingress LSR, LSRb is a color LSR and LSRc is an egress LSR. There are three paths (P1, P2 and P3) between LSRb and LSRc. Path P1 and P2 have unique characteristics with respect to p3. LSRb associates P1 and P2 with a color value C1, and P3 with a different color value C2. This color-to-path association is a local decision on LSRb, and colors are globally significant within the LDP domain.



Example of a color LSP scenario

Figure 1

LSRb signals two color LSPs towards LSRa in response to LSRc interest in color LSPs. LSRc advertises a label for a prefix for which it is an egress LSR. The advertisement includes a label mapping for the prefix and a list of colors (C1 and C2) in which LSRc is interested. LSRb matches the colors in the local color-path association with the color list that LSRc advertised. After finding a match for both C1 and C2, LSRb signals labels for two paths towards LSRa with colors C1 and C2. In addition, it programs an MPLS forwarding entry that associates color C1 with P1 and P2 as next hops, and color label C2 with P3 as next hop for the prefix advertised by LSRc. Ultimately, LSRa receives two label mappings for the prefix, one associated with C1 and the second one associated with C2.

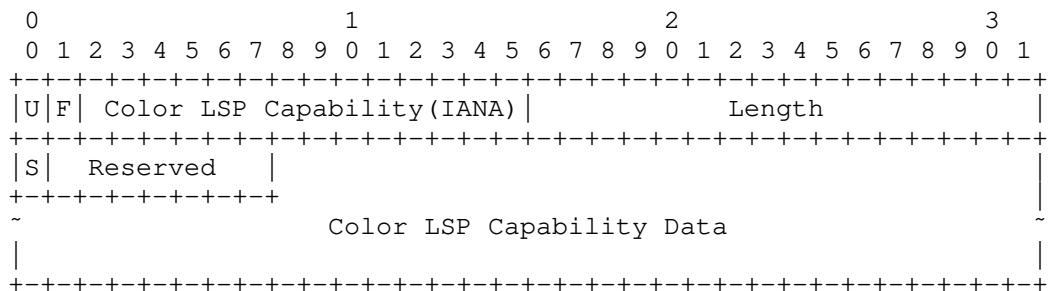
LSRa can make a local decision to forward traffic towards LSRc using the LSP with color C1 (via paths P1 and P2) or the LSP with color C2 (via P3). While both color paths overlap between LSRa and LSRb, the

latter differentiates the forwarding of color labels towards LSRC per its local policy.

LSRb may implement different policies for the signaling of color LSPs. In the scenario above, LSRb can be provisioned to advertise color labels for C1 and C2 without requiring explicit indication from LSRC of interest in those two colors. Such configuration may be desirable for interoperability with legacy egress LSRs that cannot signal a color interest. It may also be desirable in deployment scenarios where color LSPs should be signaled for all egress LSRs and there is no need to provide control for individual egress LSRs.

4. LDP Color LSP Capability

This new capability parameter allows an LSR to advertise its ability to signal a color LSP for a given FEC type (Section 2). The capability follows the format and procedures defined in [RFC5561] and may be signaled in LDP Initialization or Capability messages.



"Color LSP Capability" TLV

Figure 2

U/F-bits:

MUST be set to 1 and 0 respectively so that a receiver silently ignores this TLV if unknown, continues processing the rest of the message and does not forward the TLV if unknown.

Length:

The length (in octets) of the TLV following this length field. The value of this field is variable and is dependent on capability-specific data.

S-bit:

Set to 1 or 0 to advertise or withdraw the capability respectively as specified in [RFC5561].

Reserved:

Must be set to zero on transmission and ignored on receipt

Color LSP Capability Data:

This is capability-specific data that is defined for Color LSPs. It consists of a Typed Wildcard FEC Element [RFC5918] identifying the FEC for which this capability is being signaled. In the context of this document, the Typed Wildcard FEC element MUST correspond to one of the FEC types applicable to Color LSP as defined in Section 2. If a receiver receives this TLV with a Typed Wildcard FEC of type other than those defined in Section 2, it SHOULD silently discard the TLV and continue processing rest of the message.

In order to announce or withdraw this capability for more than one type of FEC elements, an LDP speaker MUST announce/withdraw them separately using the same "Color LSP Capability" TLV. A receiver MUST keep record of the color LSP capabilities of its peer on per-FEC basis.

For example, consider an LSR that supports color LSPs for both IPv4 Prefixes and IPv6 Prefixes. The LSR will announce these capabilities in different Capability TLVs (either as part of the same LDP Initialization/Capability message or separate message) by setting the TLV S-bit to 1 and capability data to Typed Wildcard IPv4 Prefix FEC and Typed Wildcard IPv6 Prefix FEC (as defined in [RFC5918]). The receiver will keep track of the LSR capability and note it to be Color LSP capable for IPv4-Prefix and IPv6-Prefix FEC types. Later, if the LSR withdraws its capability for one of these FEC elements, it will send a Capability TLV (in a Capability message) with S-bit set to 0 and FEC's Typed Wildcard as the capability data. On receipt of this message, the peer will update accordingly to remove the corresponding FEC from LSR's color LSP capability list.

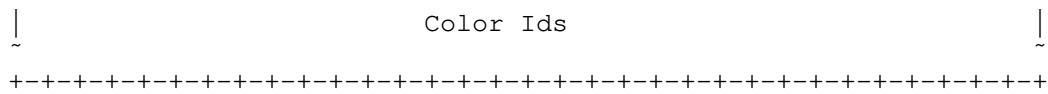
5. Color Lists**5.1. Color List TLV**

The Color List TLV is a new optional parameter in the LDP Label Mapping and Label Request messages[RFC5036]. The list includes one or more color identifiers that LSRs may use to signal interest in a forwarding preference.

```

      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
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
|U|F|           Color List (IANA)           |           Length           |
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+

```



"Color List" TLV

Figure 3

U/F-bits:

MUST be set to 1 and 1 respectively so that a receiver silently ignores this TLV if unknown, continues processing the rest of the message and forwards the TLV.

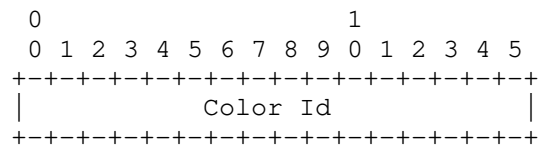
Length:

The length (in octets) of the TLV following this length field. The value of this field is variable and is dependent on number of Color Ids that follow in the TLV.

Color Ids:

List of color identifiers associated with the FEC encoded in the message.

A Color Id is a two octet field defined as follows:



"Color identifier" field

Figure 4

Color Id:

(non-zero) unsigned color identifier. Color Id 0xFFFF refers to the "wildcard color" (i.e. all colors).

5.2. Procedures

An egress LSR MAY include the Color List TLV in a Label Mapping Message if using Downstream Unsolicited mode. An LSR may include the TLV in Label Request Messages if using Downstream on Demand mode. An LSR MUST NOT include this TLV in any other LDP message except the Label Mapping and Label Request messages. LSRs SHOULD silently discard this TLV if received in other messages and continue processing the rest of the message. A Color List TLV MUST only be used in downstream signaled paths.

An LSR MAY include a Color List TLV List TLV whether the neighbor has previously advertised the LDP Color LSP Capability (Section 4) or not. The TLV MUST be forwarded to other neighbors as defined by the U/F flags. This behavior allows LSRs that do not support the Color LSP extensions to not preclude the signaling of Color LSPs if they are downstream from a Color LSR.

On receipt of a Color List TLV, a Color LSR with multiple downstream paths SHOULD match the list of color identifiers with its local association between forwarding paths and colors. At least one path MUST be defined as the "default" path. This path SHOULD be used as a second best match in the absence of an exact color match. If the Color LSR does not have an association between colors and paths or is a legacy LSR not supporting the Color LSP extensions, all paths SHOULD be treated as default paths.

6. Color Address Families

To setup LSPs corresponding to FECs under a given color scope, the applicable LDP FEC elements (Section 2) must be extended to include the color information. The Color Id becomes an attribute of such LDP FEC elements, and all FEC-Label binding operations are performed under the context of the given color.

To be able to associate a color with a FEC, we define new "color" address families as follows:

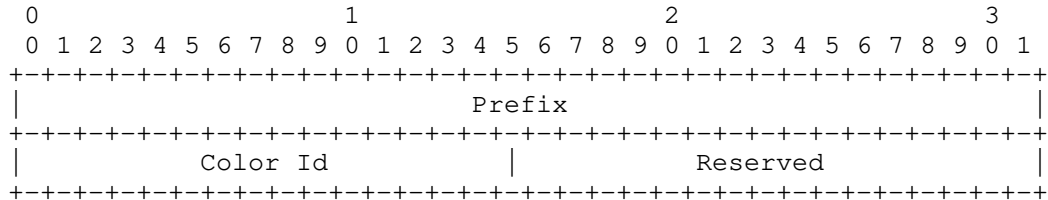
Color IP (version 4)

Color IPv6

These address families just extend the format of their base address family by including color information. The format of data associated with these new address families is described later in sections Section 6.1 and Section 6.2. The proposed new address families can be used in any LDP message and procedures defined for Color LSPs. If a receiver does not support these address families received in a message, it SHOULD send "Unknown Address Family" notification back to the sender and discard the message.

6.1. Color IP Address Family

The format of data associated with Color IP (version 4) address family is:



"Color IP (version 4)" Address Family Format

Figure 5

Prefix:

IPv4 prefix for "Color IP (version 4)" address family

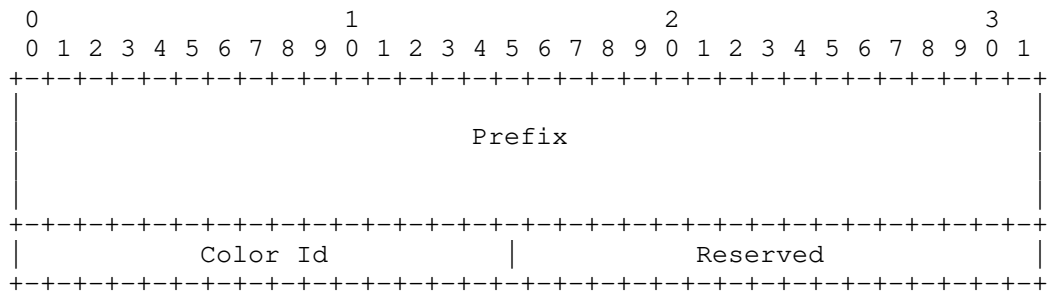
Color Id:

Color identifiers associated with IP (version 4) address

The address length for Color IP address family is 8 octets.

6.2. Color IPv6 Address Family

The format of data associated with Color IPv6 address family is:



"Color IPv6" Address Family Format

Figure 6

Prefix:

IPv6 prefix for "Color IPv6" address family

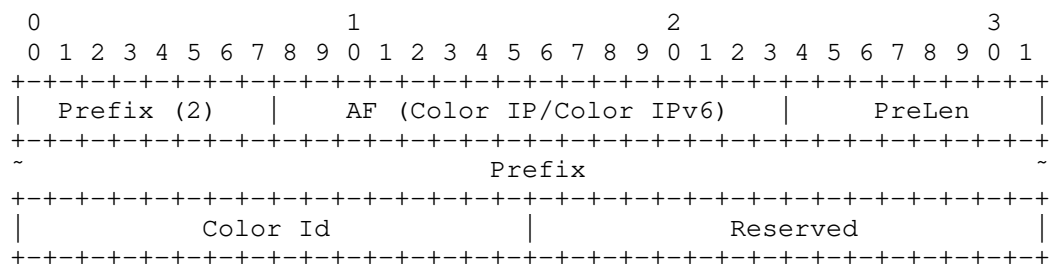
Color Id:

Color identifiers associated with IPv6 address

The address length for Color IP address family is 20 octets.

7. Color FECs

The following subsection defines the format of the LDP Color FEC elements (Section 2) as well as their Typed wildcard [RFC5918] counterparts.

7.1. Color Prefix FEC Element

"Color Prefix" FEC element

Figure 7

The definition of these fields follows Section 6 and [RFC5036]. The Prefix field can be either an IP (version 4) or an IPv6 address.

7.2. Color Multipoint FEC Elements

EDITOR NOTE: To be included in a later version.

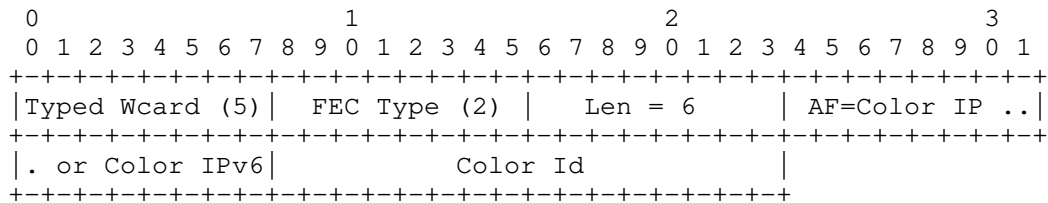
7.3. Procedures

EDITOR NOTE: To be included in a later version.

8. Other FEC-based Features**8.1. Typed Wildcard Forward Equivalence Class**

[RFC5918] extends base LDP and defines Typed Wildcard FEC Element framework. Typed Wildcard FEC element can be used in any LDP message to specify a wildcard operation for the given type of FEC. The Color LSP extensions proposed in this document do not require any extension in the procedures for Typed Wildcard FEC Element support in [RFC5918].

The encoding for a Color Prefix Typed Wildcard FEC element is as follows:



"Color Prefix Typed Wildcard" FEC element

Figure 8

The definition of these fields follows [RFC5918] and Section 5.1.

The Color Prefix Typed Wildcard FEC allows an LSR to perform wildcard FEC operations under the scope of a specific color. For example, upon local configuration of color LSP feature for a color C, an LSR may send a wildcard label request with Color Id C to learn all its labels from the peer under the scope of that color. If an LSR wishes to perform a wildcard operation that applies to all colors, it can use the "wildcard color" Color Id.

8.2. Signaling Convergence (End-of-LIB)

[RFC5919] specifies extensions and procedures that allows an LDP speaker to signal its convergence for a given FEC type towards a peer using the corresponding Typed Wildcard FEC element. Color LSP extensions for FECs do not require any change in these procedures and they apply as-is to these extended FEC elements. For instance, an LDP speaker MAY signal its LIB convergence per color (or for all colors) using a Color Prefix Typed Wildcard FEC element.

8.3. LSP Ping Extensions

8.3.1. Color Prefix FEC

[RFC4379] defines procedures to detect MPLS LSP data-plane failures via LSP ping. The section 3.2 of [RFC4379] defines Sub-Types and formats for Sub-TLVs corresponding to FECs. For "Prefix" FEC, it defines "LDP IPv4 prefix" and "LDP IPv6 prefix" sub-types and TLVs. To support LSP ping for Color LDP LSPs, this document proposes following extensions to [RFC4379]:

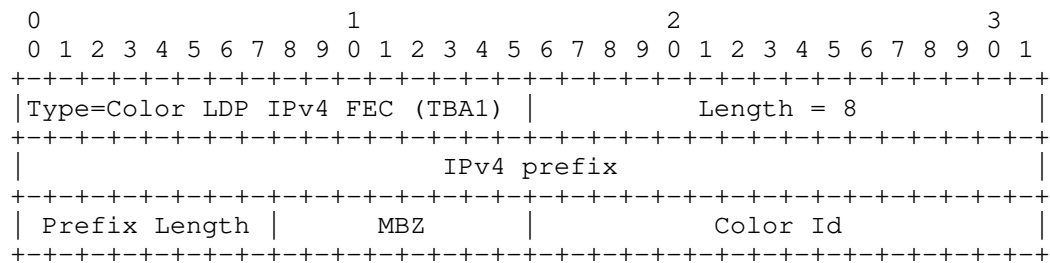
New FEC types: Color LDP IPv4 FEC, and Color LDP IPv6 FEC

New sub-types: for sub-TLVs to specify these FECs in the "Target FEC Stack" TLV of [RFC4379]

Sub-Type	Length	Value Field
TBA1	5	Color LDP IPv4 prefix
TBA2	17	Color LDP IPv6 prefix

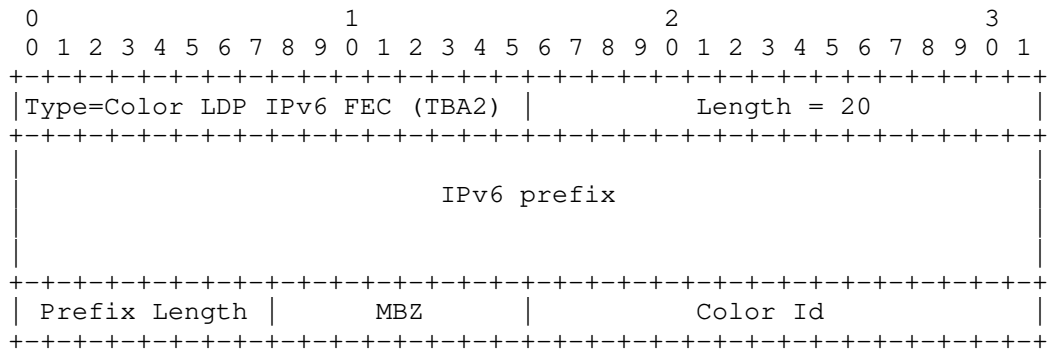
The format of these new FEC types is defined as an extension to the format of LDP IPv4 prefix and LDP IPv6 prefix sub-TLV by adding color information.

The encoding for a Color LDP IP (version 4) and IPv6 FEC sub-TLVs is as follows:



"Color LDP IP (version 4)" FEC sub-TLV for LSP ping

Figure 9



"Color LDP IPv6" FEC sub-TLV for LSP ping

Figure 10

The Color Id value MUST NOT be the "Wildcard Color".

8.3.2. Color Multipoint FEC

EDITOR NOTE: To be included in a later version.

9. IANA Considerations

This document defines the following new LDP extensions:

"Color LSP Capability" (codepoint to be allocated from LDP registry "TLV Type Name Space")

Color List TLV (codepoint to be allocated from LDP registry "TLV Type Name Space")

Color Address Families (from registry "Address Family Numbers")

Color IP (version 4)

Color IPv6

New Sub-TLV Types under TLV type 1 (Target FEC Stack) from "Multi-Protocol Label Switching (MPLS) Label Switched Paths (LSPs) Ping Parameters" registry, and "TLVs and sub-TLVs" sub-registry.

Sub-Type	Value Field	
TBA1	5	Color LDP IPv4 prefix
TBA2	17	Color LDP IPv6 prefix

IANA is requested to assign the LDP capability code point and the type values of these TLVs.

10. Security Considerations

The MPLS security framework [RFC5920] and the security considerations in the LDP specification [RFC5036] apply to this document.

11. References

11.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC4379] Kompella, K. and G. Swallow, "Detecting Multi-Protocol Label Switched (MPLS) Data Plane Failures", RFC 4379, February 2006.
- [RFC5036] Andersson, L., Minei, I., and B. Thomas, "LDP Specification", RFC 5036, October 2007.
- [RFC5561] Thomas, B., Raza, K., Aggarwal, S., Aggarwal, R., and JL. Le Roux, "LDP Capabilities", RFC 5561, July 2009.
- [RFC5918] Asati, R., Minei, I., and B. Thomas, "Label Distribution Protocol (LDP) 'Typed Wildcard' Forward Equivalence Class (FEC)", RFC 5918, August 2010.
- [RFC5919] Asati, R., Mohapatra, P., Chen, E., and B. Thomas, "Signaling LDP Label Advertisement Completion", RFC 5919, August 2010.
- [RFC5920] Fang, L., "Security Framework for MPLS and GMPLS Networks", RFC 5920, July 2010.
- [RFC6388] Wijnands, IJ., Minei, I., Kompella, K., and B. Thomas, "Label Distribution Protocol Extensions for Point-to-Multipoint and Multipoint-to-Multipoint Label Switched Paths", RFC 6388, November 2011.

11.2. Informative References

- [RFC3031] Rosen, E., Viswanathan, A., and R. Callon, "Multiprotocol Label Switching Architecture", RFC 3031, January 2001.
- [RFC5920] Fang, L., "Security Framework for MPLS and GMPLS Networks", RFC 5920, July 2010.

Authors' Addresses

Santiago Alvarez
Cisco Systems, Inc.
170 West Tasman Drive
San Jose, CA 95134
USA

Email: saalvare@cisco.com

Kamran Raza
Cisco Systems, Inc.
2000 Innovation Drive
Kanata, ON K2K-3E8
Canada

Email: skraza@cisco.com

Sami Boutros
Cisco Systems, Inc.
170 West Tasman Drive
San Jose, CA 95134
USA

Email: sboutros@cisco.com

Network Working Group
Internet-Draft
Updates: 4379 (if approved)
Intended status: Standards Track
Expires: January 16, 2014

L. Andersson
Huawei
July 15, 2013

Updates to RFC 4379 IANA section
draft-andersson-mpls-lsp-ping-upd-00

Abstract

The MultiProtocol Label Switching (MPLS) protocol for detecting Label Switched Path failures (LSP Ping), as defined in RFC 4379 and several extensions, are widely deployed and very popular.

The IANA section of RFC4379 lack in clarity and need to be updated.

Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [RFC2119].

Status of This Memo

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

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

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

This Internet-Draft will expire on January 16, 2014.

Copyright Notice

Copyright (c) 2013 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	2
2. Update of the RCF 4379 IANA section	2
2.1. Error codes for unrecognized TLV and sub-TLV types	3
3. IANA Considerations	3
4. Security Considerations	3
5. Acknowledgements	3
6. Normative References	3
Author's Address	4

1. Introduction

This document updates the IANA section of RFC 4379 [RFC4379] for LSP Ping Parameters.

2. Update of the RCF 4379 IANA section

The first 3 paragraphs of sub-section 7.2. "TLVs" in the IANA section of RFC 4379 is considered unclear and is therefore now replaced by the following text:

The IANA has created and will maintain a registry for the Type field of top-level TLVs and per-TLV registries for any TLVs which have associated sub-TLVs. Note the meaning of a sub-TLV is scoped by the TLV. The number spaces for the sub-TLVs of various TLVs are independent.

However, it is under some conditions allowable for a new TLV to re-use sub-TLVs of another TLV. In this case where all sub-TLVs are re-used and no unique sub-TLVs are defined for the TLV re-uses the sub-TLVs no actual sub-TLV registry is created for the new TLV. Rather an entry is made where the registry would have appeared with a note saying "Uses the sub-TLVs registered under TLV x", where x is the other TLV. Note that the implication here is that all future sub-TLVs of the other TLV apply, as well as those currently defined.

The valid range for TLV registry is 0-65535 and this is also default for each of the sub-TLV registries. For all these registries, assignments in the range 0-16383 and 32768-49161 are made via

Standards Action as defined in [IANA]; assignments in the range 16384-31743 and 49162-64511 are made via "Specification Required" as defined above; values in the range 31744-32767 and 64512-65535 are for Vendor Private Use, and MUST NOT be allocated.

However, a new TLV type might specify the allocation policies for its own sub-TLVs.

2.1. Error codes for unrecognized TLV and sub-TLV types

For TLV and sub-TLV types that uses the default allocation ranges defined above the rules for when error messages defined below applies. TLVs that defines there own sub-TLV ranges, does also need to define their own rules for when error messages are returned.

TLV and sub-TLV types less than 32768 (i.e., with the high-order bit equal to 0) are mandatory TLVs that MUST either be supported by an implementation or result in the return code of 2 ("One or more of the TLVs was not understood") being sent in the echo response.

TLV and sub-TLV types greater than or equal to 32768 (i.e., with the high-order bit equal to 1) are optional TLVs that SHOULD be ignored if the implementation does not understand or support them.

If a TLV or sub-TLV has a Type that falls in the range for Vendor Private Use, the Length MUST be at least 4, and the first four octets MUST be that vendor's SMI Private Enterprise Number, in network octet order. The rest of the Value field is private to the vendor.

3. IANA Considerations

There are no requests for IANA actions in this document.

4. Security Considerations

This document is about updating the IANA section of RFC 4379 and does not add any new security issues as compared to the the original RFC 4379.

5. Acknowledgements

6. Normative References

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

[RFC4379] Kompella, K. and G. Swallow, "Detecting Multi-Protocol Label Switched (MPLS) Data Plane Failures", RFC 4379, February 2006.

Author's Address

Loa Andersson
Huawei

Email: loa@mail01.huawei.com

Network Working Group
Internet-Draft
Updates: 6374 (if approved)
Intended status: Standards Track
Expires: January 15, 2014

L. Andersson
Huawei
July 14, 2013

Moving Generic Associated Channel registries to a new name space
draft-andersson-mpls-moving-iana-registries-00

Abstract

When RFC 6374 "Packet Loss and Delay Measurement for MPLS Networks" were developed, the code points allocated by this RFC were mistakenly placed within the Multi-Protocol Label Switching (MPLS) Label Switched Paths (LSPs) Ping Parameter registry. This document creates a new dedicated name space for Generic Associated Channel code points and move them to a name space their own.

Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [RFC2119].

Status of This Memo

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

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

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

This Internet-Draft will expire on January 15, 2014.

Copyright Notice

Copyright (c) 2013 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	2
2. Moving GACH parameters out of LSP Ping parameter registry . .	2
3. IANA Considerations	3
3.1. Moving Loss/Delay registries	3
4. Security Considerations	4
5. Acknowledgements	4
6. References	4
6.1. Normative References	4
6.2. Informative References	4
Author's Address	5

1. Introduction

RFC 6374 [RFC6374] specify protocols for delay and loss measurements. The protocols run over the Generic Associated Channel (GACH). There are four registries with code points for these protocols. These registries were mistakenly allocated within the LSP Ping Parameters namespace. This document now creates a new GACH name space and move the Loss and Delay parameter registries to the new registry.

2. Moving GACH parameters out of LSP Ping parameter registry

In the "Multi-Protocol Label Switching (MPLS) Label Switched Paths (LSPs) Ping Parameters" name space there are four registries that are for using the MPLS Generic Associated Channel [RFC5586] for MPLS Loss and Delay measurements [RFC6374].

The registries are:

- o Measurement Timestamp Type
- o Loss/Delay Measurement Control Code: Query Codes
- o Loss/Delay Measurement Control Code: Response Codes
- o MPLS Loss/Delay Measurement TLV Object

These registries are now moved into a new name space under the "Multiprotocol Label Switching Architecture" (MPLS) heading called "MPLS Generic Associated Channel Parameters".

This is an update to RFC6374.

The sub-section 9.2 in the IANA section of RFC6374 has a sentence that read:

IANA has created a new "Measurement Timestamp Type" registry, with format and initial allocations as follows:

This sentence is now changed to read:

IANA has created a new "Measurement Timestamp Type" registry in the "MPLS Generic Associated Channel Parameters" name space, with format and initial allocations as follows:

The sub-section 9.3 in the IANA section of RFC6374 has a sentence that read:

IANA has created a new "MPLS Loss/Delay Measurement Control Code" registry.

This sentence is now changed to read:

IANA has created a new "MPLS Loss/Delay Measurement Control Code" registry in the "MPLS Generic Associated Channel Parameters" name space, with format and initial allocations as follows:

The sub-section 9.4 in the IANA section of RFC6374 has a sentence that read:

IANA has created a new "MPLS Loss/Delay Measurement TLV Object" registry, with format and initial allocations as follows:

This sentence is now changed to read:

IANA has created a new "MPLS Loss/Delay Measurement TLV Object" registry in the "MPLS Generic Associated Channel Parameters", with format and initial allocations as follows:

3. IANA Considerations

IANA is requested to take the actions listed below.

3.1. Moving Loss/Delay registries

1. To create a new name space called "MPLS Generic Associated Channel Parameters" under the "MPLS Architecture" heading.
2. To move the "Measurement Timestamp Type", the "Loss/Delay Measurement Control Code: Query Codes", the "Loss/Delay Measurement Control Code: Response Codes" and the "MPLS Loss/Delay Measurement TLV Object" to the newly created "MPLS Generic Associated Channel Parameters" name space.
3. Not to change any of the values previously assigned in the registries that are moved.
4. The registries should be moved without any traces left in the "Multi-Protocol Label Switching (MPLS) Label Switched Paths (LSPs) Ping Parameters" name space.

4. Security Considerations

This document is about updating the IANA LSP Ping TLV and sub-TLV registries and it does not add any new security issues as compared to the RFC that specifies the protocol.

5. Acknowledgements

6. References

6.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC4379] Kompella, K. and G. Swallow, "Detecting Multi-Protocol Label Switched (MPLS) Data Plane Failures", RFC 4379, February 2006.
- [RFC5586] Bocci, M., Vigoureux, M., and S. Bryant, "MPLS Generic Associated Channel", RFC 5586, June 2009.
- [RFC6374] Frost, D. and S. Bryant, "Packet Loss and Delay Measurement for MPLS Networks", RFC 6374, September 2011.

6.2. Informative References

- [IANA-LSP-Ping]
 , "LSP Ping Parameters", , <<http://www.iana.org/assignments/mpls-lsp-ping-parameters/mpls-lsp-ping-parameters.xml>>.

Author's Address

Loa Andersson
Huawei

Email: loa@mail01.huawei.com

MPLS Working Group
Internet-Draft
Intended status: Standards Track
Expires: January 13, 2014

A. Atlas
K. Tiruveedhula
Juniper Networks
J. Tantsura
Ericsson
IJ. Wijnands
Cisco Systems, Inc.
July 12, 2013

LDP Extensions to Support Maximally Redundant Trees
draft-atlas-mpls-ldp-mrt-00

Abstract

This document specifies extensions to LDP to support the creation of label-switched paths for Maximally Redundant Trees (MRT). A prime use of MRTs is for unicast and multicast IP/LDP Fast-Reroute (MRT-FRR).

The sole protocol extension to LDP is simply the ability to advertise an MRT Capability. This document describes that extension and the associated behavior expected for LSRs and LERs advertising the MRT Capability.

MRT-FRR uses LDP multi-topology extensions and requires three different multi-topology IDs to be allocated from the LDP MT-ID space.

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 January 13, 2014.

Copyright Notice

Copyright (c) 2013 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	2
2. Requirements Language	3
3. Terminology	3
4. Overview of LDP Signaling Extensions for MRT	4
4.1. MRT Capability Advertisement	5
4.2. Behavior Related to the Rainbow MRT MT-ID	6
4.3. MRT-Blue and MRT-Red FECs	6
5. LDP MRT FEC Advertisements	7
5.1. Downstream Unsolicited Mode	7
5.2. Downstream On Demand Mode	7
5.3. Inter-Area	8
6. Security Considerations	8
7. IANA Considerations	8
8. Acknowledgements	9
9. References	9
9.1. Normative References	9
9.2. Informative References	9
Authors' Addresses	10

1. Introduction

This document describes the LDP signaling extension and associated behavior necessary to support the architecture that defines how IP/LDP Fast-Reroute can use MRTs [I-D.ietf-rtgwg-mrt-frr-architecture]. It is necessary to read the architecture in [I-D.ietf-rtgwg-mrt-frr-architecture] to understand how and why the LDP extensions for behavior are needed.

At least one common standardized algorithm, such as the lowpoint algorithm explained and fully documented in [I-D.enyedi-rtgwg-mrt-frr-algorithm], is required so that the routers supporting MRT computation consistently compute the same MRTs. LDP depends on the IGP to compute the MRTs and alternates; extensions to OSPF are defined in [I-D.atlas-ospf-mrt].

MRT can also be used to protect multicast traffic via either global protection or local protection. [I-D.atlas-rtgwg-mrt-mc-arch] An MRT path can be used to provide node-protection for mLDP traffic via the mechanisms described in [I-D.wijnands-mpls-mldp-node-protection]; an MRT path can also be use to provide link protection for mLDP traffic.

For each destination, IP/LDP Fast-Reroute with MRT (MRT-FRR) creates two alternate destination-based trees separate from the primary next-hop forwarding used during stable operation. LDP uses the multi-topology extensions [I-D.ietf-mpls-ldp-multi-topology] to signal FECs for these two new forwarding topologies, known as MRT-Blue and MRT-Red.

In order to create MRT paths and support IP/LDP Fast-Reroute, a new capability extension is needed for LDP. An LDP implementation supporting MRT must also follow the described rules for originating and managing FECs related to MRT, as indicated by their multi-topology ID. Network reconvergence is described in [I-D.ietf-rtgwg-mrt-frr-architecture] and the worst-case network convergence time can be flooded via the extension in Section 7 of [I-D.atlas-ospf-mrt].

IP/LDP Fast-Reroute using MRTs can provide 100% coverage for link and node failures in an arbitrary network topology where the failure doesn't split the network. It can also be deployed incrementally; an MRT Island is formed of connected supporting routers and the MRTs are computed inside that island.

2. 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]

3. Terminology

For ease of reading, some of the terminology defined in [I-D.ietf-rtgwg-mrt-frr-architecture] is repeated here.

Redundant Trees (RT): A pair of trees where the path from any node X to the root R along the first tree is node-disjoint with the

path from the same node X to the root along the second tree. These can be computed in 2-connected graphs.

Maximally Redundant Trees (MRT): A pair of trees where the path from any node X to the root R along the first tree and the path from the same node X to the root along the second tree share the minimum number of nodes and the minimum number of links. Each such shared node is a cut-vertex. Any shared links are cut-links. Any RT is an MRT but many MRTs are not RTs. The two MRTs are referred to as MRT-Blue and MRT-Red.

MRT Island: From the computing router, the set of routers that support a particular MRT profile and are connected via MRT-eligible links.

MRT-Red: MRT-Red is used to describe one of the two MRTs; it is used to describe the associated forwarding topology and MT-ID. Specifically, MRT-Red is the decreasing MRT where links in the GADAG are taken in the direction from a higher topologically ordered node to a lower one.

MRT-Blue: MRT-Blue is used to describe one of the two MRTs; it is used to describe the associated forwarding topology and MT-ID. Specifically, MRT-Blue is the increasing MRT where links in the GADAG are taken in the direction from a lower topologically ordered node to a higher one.

Rainbow MRT: It is useful to have an MT-ID that refers to the multiple MRT topologies and to the default topology. This is referred to as the Rainbow MRT MT-ID and is used by LDP to reduce signaling and permit the same label to always be advertised to all peers for the same (MT-ID, Prefix).

4. Overview of LDP Signaling Extensions for MRT

Routers need to know which of their neighbors support MRT. Supporting MRT indicates several different aspects of behavior, as listed below.

1. Support for Multi-Topology (MT) - this MAY also be indicated via the Multi-Capability MT Capability [I-D.ietf-mpls-ldp-multi-topology].
2. Understand the Rainbow MRT MT-ID and apply the associated labels to all relevant MT-IDs.
3. Advertise the Rainbow MRT MT-ID to the appropriate neighbors for the associated prefix.

4. If acting as egress for a prefix in the default topology, also advertise and act as egress for the same prefix in MRT-Red and MRT-Blue.
5. For a FEC learned from a neighbor that does not support MRT, originate FECS for MRT-Red and MRT-Blue with the same prefix.

4.1. MRT Capability Advertisement

It is not possible to support MRT without supporting the LDP multi-topology extensions, but it is possible that the only use of the multi-topology extensions is for MRT. In that case, a router MAY not negotiate the multi-topology capability and only negotiate the MRT Capability with its LDP peer. Negotiation of the MT capability is not required with negotiation of the MRT capability.

[EDITOR NOTE: How do we deal with different abilities for IPv4 and IPv6? The MT capability has the Wildcard FEC to indicate this. Do we just assume??]

A new MRT Capability Parameter TLV is defined, which is defined in accordance with LDP Capability definition guidelines[RFC5561].

The LDP MRT capability can be advertised during the LDP session initialization or after the LDP session is established. Advertisement of the MRT capability indicates support of the procedures for establishing the MRT-Blue and MRT-Red LSP paths detailed in this document. If the peer has not advertised the corresponding capability, then it indicates that LSR is not capable of supporting MRT procedures.

The following is the format of the MRT Capability Parameter.

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																																
U F MRT Capability (IANA)																Length (= 1)																																															
S Reserved																																																															

MRT Capability TLV Format

Where:

U- and F-bits: MUST be 1 and 0, respectively, as per Section 3. (Signaling Extensions) of LDP Capabilities [RFC5561].

MRT Capability: Capability TLV type (IANA assigned)

S-bit: MUST be 1 if used in LDP "Initialization" message. MAY be set to 0 or 1 in dynamic "Capability" message to advertise or withdraw the capability respectively.

Length: The length (in octets) of TLV. Its value is 1.

4.2. Behavior Related to the Rainbow MRT MT-ID

In Section 9 of [I-D.ietf-rtgwg-mrt-frr-architecture], the need to advertise different MPLS labels to different neighbors for the same FEC is described. This can be shortly summarized as either advertising MRT MT-ID differentiated labels to a neighbor or just advertising the same MPLS label for the default topology, for MRT-Red and MRT-Blue. MRT-supporting neighbors in the same domain as the default SPT next-hop get the differentiated MPLS labels; all other neighbors do not.

A second use for the Rainbow MRT MT-ID is for an egress LER to send the Rainbow MRT MT-ID with an IMPLICIT_NULL label to indicate penultimate-hop-popping for all three types of FECs (IP Prefix FEC, MRT-Blue MT-IP Prefix FEC, and MRT-Red MT-IP Prefix FEC).

An LSR advertising the MRT capability MUST recognize the Rainbow MRT MT-ID and associate the advertised label with the specific prefix for the default topology (MT-ID 0) and with the MRT-Red and MRT-Blue MT-IDs associated with all MRT Profiles that advertise LDP as the forwarding mechanism.

An LSR is RECOMMENDED to use the Rainbow MRT MT-ID to reduce the amount of state and signaling required.

As described in [I-D.ietf-rtgwg-mrt-frr-architecture], the recommended experimental value for the Rainbow MRT MT-ID is 3999. The final value will be assigned by IANA and allocated from the LDP MT-ID space.

4.3. MRT-Blue and MRT-Red FECs

To provide MRT support in LDP, the MT Prefix FEC is used. For the default MRT Profile, an MRT-Blue FEC uses the MRT-Blue MT-ID value TBD3 allocated by IANA; for experimental purposes, the value 3998 is suggested. For the default MRT Profile, an MRT-Red FEC uses the MRT-Red MT-ID value TBD2 allocated by IANA; for experimental purposes, the value 3997 is suggested.

The MT Prefix FEC encoding is defined in [I-D.ietf-mpls-ldp-multi-topology] and is used without alternation for signaling MRT-Blue, MRT-Red and Rainbow MRT FECs.

5. LDP MRT FEC Advertisements

This sections describes how and when labels for MRT-Red and MRT-Blue FECs are advertised. The associated LSPs must be created before any failure occurs.

5.1. Downstream Unsolicited Mode

If the upstream session is negotiated with the MRT capability, the Egress LER advertises via a Rainbow MRT FEC an allocated MPLS label; this may be Explicit Null, Implicit Null, or another value.

Based on the MRT algorithm [I-D.enyedi-rtgwg-mrt-frr-algorithm], the IGP computes the MRT-Red and MRT-Blue disjoint paths at Ingress and Transit LSRs. Once the IGP computes the MRT-Red and MRT-Blue next-hops, LDP will advertise the Label Mapping for the MRT-Blue and MRT-Red FECs. If a label is received from a downstream LSR for an MRT-Red or MRT-Blue FEC where the downstream LSR is capable of MRT, the MRT-Red FEC or MRT-Blue FEC label is swapped according to the received downstream label. An LSR may also choose to use the MRT-Red or MRT-Blue path as an alternative for doing fast-reroute for the local traffic.

When a downstream router is not capable of MRT, the LSR is an MRT Island Border Router (IBR) and SHOULD advertise Label Bindings for the MRT-Red FEC and MRT-Blue FEC as well as the associated normal topology. The normal topology's primary next-hops will be used to forward traffic received for the MRT-Red FEC or the MRT-Blue FEC where the FEC's destination is outside the MRT Island. This functionality is critical for partial deployment scenarios.

5.2. Downstream On Demand Mode

After the IGP computes the MRT-Red and MRT-Blue paths, the IGP MAY also decide to use either the MRT-Red or MRT-Blue path as a fast-reroute alternate for the particular FEC. If so, then when in Downstream On Demand Mode, the LSR sends a Label Request for either the MRT-Red or MRT-Blue FEC to the downstream LSR. The downstream LSR responds by either sending a Label Mapping if available or by sending a Label Request to its downstream LSR. Once a Label Mapping is received, the associated label may be used as a fast-reroute alternative to forward IP and LDP traffic.

A Label Mapping may be available in the following circumstances:

- o The LSR is acting as Egress
- o A Label Mapping was already received from its downstream router
- o A Label Mapping for the default topology FEC was received and the downstream router is not capable of MRT or is in a different MRT Island.

5.3. Inter-Area

As discussed in Section 4.2, the Rainbow MRT FEC is defined to facilitate signaling the same label for multiple topologies. Section 9 of [I-D.ietf-rtgwg-mrt-frr-architecture] recommends that traffic leaving an OSPF area or IS-IS level SHOULD use the default topology's shortest-path-tree next-hops instead of remaining on the MRT-Red or MRT-Blue paths. If an LDP peer is in the same OSPF area or IS-IS level as the primary next-hop, then LDP SHOULD advertise different label values for a given set of MRT-Red FEC, MRT-Blue FEC, and FEC, unless Explicit-Null or Implicit-Null is appropriate. If an LDP peer is in a different OSPF area or IS-IS level from the primary next-hop, then LDP SHOULD either advertise the same label value for the given set of MRT-Red FEC, MRT-Blue FEC, and FEC or advertise a single label for the Rainbow MRT FEC, whose behavior is defined in Section 4.2.

6. Security Considerations

This LDP extension is not believed to introduce new security concerns. It relies upon the security architecture already provided for LDP.

7. IANA Considerations

New LDP Capability TLV: "MRT Capability" TLV (requested code point: TBA from LDP registry "TLV Type Name Space"). For interoperable experimental purposes, the value of ... is suggested.

Allocations from the "LDP Multi-Topology (MT) ID Name Space" [I-D.ietf-mpls-ldp-multi-topology] under "LDP Parameter" namespace:

- o Rainbow MRT MT-ID: TBD1
- o default Profile MRT-Red MT-ID: TBD2 - requested under 4096 so it can also be signaled in PIM
- o default Profile MRT-Blue MT-ID: TBD3 - requested under 4096 so it can also be signaled in PIM

For interoperable experiments, the following values are suggested for experimentation: Rainbow MRT MT-ID 3999, default MRT Profile MRT-Blue MT-ID 3998, default MRT Profile MRT-Red MT-ID 3997. The MT-IDs are taken from the 3996-4096 range, which IS-IS defines as for private use, and which [I-D.ietf-mpls-ldp-multi-topology] does not specify as reserved (and MPLS list email suggests that range may be reserved for private use mapping from the IS-IS space).

8. Acknowledgements

The authors would like to thank Ross Callon for his suggestions.

9. References

9.1. Normative References

- [I-D.ietf-mpls-ldp-multi-topology]
Zhao, Q., Fang, L., Zhou, C., Li, L., and K. Raza, "LDP Extensions for Multi Topology Routing", draft-ietf-mpls-ldp-multi-topology-08 (work in progress), May 2013.
- [I-D.ietf-rtgwg-mrt-frr-architecture]
Atlas, A., Kebler, R., Envedi, G., Csaszar, A., Tantsura, J., Konstantynowicz, M., and R. White, "An Architecture for IP/LDP Fast-Reroute Using Maximally Redundant Trees", draft-ietf-rtgwg-mrt-frr-architecture-03 (work in progress), July 2013.
- [RFC5561] Thomas, B., Raza, K., Aggarwal, S., Aggarwal, R., and JL. Le Roux, "LDP Capabilities", RFC 5561, July 2009.

9.2. Informative References

- [I-D.atlas-ospf-mrt]
Atlas, A., Hegde, S., Chris, C., and J. Tantsura, "OSPF Extensions to Support Maximally Redundant Trees", draft-atlas-ospf-mrt-00 (work in progress), July 2013.
- [I-D.atlas-rtgwg-mrt-mc-arch]
Atlas, A., Kebler, R., Wijnands, I., Csaszar, A., and G. Envedi, "An Architecture for Multicast Protection Using Maximally Redundant Trees", draft-atlas-rtgwg-mrt-mc-arch-02 (work in progress), July 2013.
- [I-D.envedi-rtgwg-mrt-frr-algorithm]

Atlas, A., Envedi, G., Csaszar, A., Gopalan, A., and C. Bowers, "Algorithms for computing Maximally Redundant Trees for IP/LDP Fast- Reroute", draft-envedi-rtgwg-mrt-frr-algorithm-03 (work in progress), July 2013.

[I-D.wijnands-mpls-mldp-node-protection]

Wijnands, I., Rosen, E., Raza, K., Tantsura, J., Atlas, A., and Q. Zhao, "mLDP Node Protection", draft-wijnands-mpls-mldp-node-protection-04 (work in progress), June 2013.

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

[RFC4915] Psena, P., Mirtorabi, S., Roy, A., Nguyen, L., and P. Pillay-Esnault, "Multi-Topology (MT) Routing in OSPF", RFC 4915, June 2007.

[RFC5715] Shand, M. and S. Bryant, "A Framework for Loop-Free Convergence", RFC 5715, January 2010.

Authors' Addresses

Alia Atlas
Juniper Networks
10 Technology Park Drive
Westford, MA 01886
USA

Email: akatlas@juniper.net

Kishore Tiruveedhula
Juniper Networks
10 Technology Park Drive
Westford, MA 01886
USA

Email: kishoret@juniper.net

Jeff Tantsura
Ericsson
300 Holger Way
San Jose, CA 95134
USA

Email: jeff.tantsura@ericsson.com

IJsbrand Wijnands
Cisco Systems, Inc.

Email: ice@cisco.com

MPLS Working Group
Internet-Draft
Intended status: Standards Track
Expires: September 12, 2013

T. Cheung
ETRI
A. D'Alessandro
Telecom Italia
H. van Helvoort
Huawei Technologies
March 11, 2013

PSC protocol updates for non-revertive operation
draft-cdh-mpls-tp-psc-non-revertive-00.txt

Abstract

This document contains the updates to [RFC6378], "MPLS Transport Profile (MPLS-TP) Linear Protection" to change non-revertive operation to be aligned with the behavior defined in [RFC4427] and in an effort to satisfy the ITU-T's protection switching requirements. An operator command, Manual Switch to Working (MS-W) is also included to revert traffic to the working path in non-revertive operation.

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 September 12, 2013.

Copyright Notice

Copyright (c) 2013 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. Motivation for adding MS-W	3
1.2. Behavior of MS-P and MS-W	4
1.3. Equal priority resolution	4
2. Conventions Used in This Document	4
3. Acronyms	4
4. Updates to the PSC RFC	4
4.1. Updates to Section 2.1. Acronyms	5
4.2. Updates to Section 3.1. Local Request Logic	5
4.3. Updates to Section 3.2. Remote Requests	6
4.4. Updates to Section 3.6. PSC Control States	7
4.5. Updates to Section 4.2.2. PSC Request Field	7
4.6. Updates to Section 4.3.2. Priority of Inputs	8
4.7. Updates to Section 4.3.3.1. Normal State	10
4.8. Updates to Section 4.3.3.2. Unavailable State	11
4.9. Updates to Section 4.3.3.3. Protecting Administrative State	12
4.10. Updates to Section 4.3.3.4. Protecting Failure State	16
4.11. Updates to Section 4.3.3.5. Wait-to-Restore State	16
4.12. Updates to Section 4.3.3.6. Do-not-Revert State	18
4.13. Updates to Appendix A. PSC State Machine Tables	20
5. Security considerations	22
6. IANA considerations	23
7. Acknowledgements	23
8. References	23
8.1. Normative References	23
8.2. Informative References	23
Authors' Addresses	23

1. Introduction

Non-revertive mode of protection switching is defined in [RFC4427]. In this mode, the traffic does not return to the working path when switch-over requests are terminated.

However, PSC protocol defined in [RFC6378] supports this operation only when recovering from a defect condition, but does not operate as non-revertive when an operator's switch-over command such as Forced Switch or Manual Switch is cleared. To be aligned with legacy transport network behavior and [RFC4427], a node should go into the Do-not-Revert (DNR) state not only when a failure condition on a working path is cleared but also when an operator command requesting switch-over is cleared.

Changing the non-revertive operation introduces necessity of a new operator command to revert traffic to the working path when in DNR state. Moreover, according to Section 4.3.3.6. Do-not-Revert State in [RFC6378], "to revert back to Normal state, the administrator SHALL issue a Lockout of protection command followed by a Clear command." This requirement introduces the potential risk of an unprotected situation while the Lockout of protection is in effect. Manual Switch-over for recovery LSP/span command, defined in [RFC4427] and also defined in [RFC5654], Requirement 83, as one of the mandatory external commands, should be used for this purpose, but is not included in [RFC6378].

It should be noted that the missing of this command from [RFC6378] is identified in the ITU-T's liaison statements [LIAISON1205] and [LIAISON1234].

This document contains the updates to [RFC6378] to change non-revertive operation to be aligned with the behavior defined in [RFC4427] and to meet the ITU-T's protection switching requirements, and add a new operator command, Manual Switch to Working (MS-W) to avoid the potential problem with the Lockout of protection command when the DNR should be cleared.

1.1. Motivation for adding MS-W

Most of the operational interventions on working paths are executed after operating a "Manual switch-over for normal traffic" switch command that switches the normal traffic from the working path to the protection path. This command will keep the traffic on the protection path unless a "Manual switch-over for recovery LSP/span" command is issued that switches the normal traffic back to the working path. Using Lockout of protection command as currently suggested in [RFC6378] may cause, in some circumstances, traffic

loss.

1.2. Behavior of MS-P and MS-W

The MS-P and MS-W commands SHALL have the same priority. If one of these commands is already issued, and the other command is issued afterwards, it SHALL be ignored. If two LERs are requesting opposite operations simultaneously, i.e. one LER is sending MS-P while the other LER is sending MS-W, the MS-W SHALL be considered to have a higher priority than MS-P, and MS-P SHALL be ignored.

This behavior is described in Section 4.2 that proposes updates to Section 3.1 "Local Request Logic" of [RFC6378].

1.3. Equal priority resolution

[RFC6378] defines only one rule for equal priority condition in Section 4.3.2 as "The remote message from the far-end LER is assigned a priority just below the similar local input." In order to support the manual switch behavior described in Section 1.2, additional rules for equal priority resolution are required, and are described in Section 4.6 that proposes updates to Section 4.3.2. "Priority of Inputs" of [RFC6378].

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. Acronyms

This draft uses the following acronyms:

MPLS-TP	Transport Profile for MPLS
MS	Manual Switch
MS-P	Manual Switch to Protection
MS-W	Manual Switch to Working
PSC	Protection State Coordination Protocol

4. Updates to the PSC RFC

This section describes the changes required to change non-revertive operation and add Manual Switch to Working operator command in the PSC protocol defined in [RFC6378].

The term "Manual Switch" and its acronym "MS" used in [RFC6378] are replaced respectively by "Manual Switch to Protection" and "MS-P" by this document to avoid confusion with "Manual Switch to Working" and its acronym "MS-W".

Also, the term "Protecting administrative state" used in [RFC6378] is replaced by "Switching administrative state" by this document to include the case where traffic is switched back to the working path by administrative Manual Switch to Working command.

4.1. Updates to Section 2.1. Acronyms

Replace the following bullet item:

MS Manual Switch

With:

MS-P Manual Switch to Protection

MS-W Manual Switch to Working

4.2. Updates to Section 3.1. Local Request Logic

Replace the following text in the bullet item for operator command:

The commands Forced Switch, Manual Switch, Clear, Lockout of protection (defined in [RFC4427] as Forced switch-over, Manual switch-over, Clear, and Lockout of recovery LSP/span, respectively) MUST be supported.

With:

The commands Forced Switch, Manual Switch to Protection, Manual Switch to Working, Clear, Lockout of protection (defined in [RFC4427] as Forced switch-over for normal traffic, Manual switch-over for normal traffic, Manual switch-over for recovery LSP/span, Clear and Lockout of recovery LSP/span, respectively) MUST be supported.

Replace the following bullet item in the local request list:

- o Manual Switch (MS) - if the operator requested that traffic be switched from the working path to the protection path. This is only relevant if there is no currently active fault condition or operator command.

With:

- o Manual Switch to Protection (MS-P) - if the operator requested that traffic be switched from the working path to the protection path. This is only relevant if there is no currently active fault condition or operator command.
- o Manual Switch to Working (MS-W) - if the operator requested that traffic be switched from the protection path to the working path. This is only relevant if there is no currently active fault condition or operator command.

Add the following text above the last paragraph:

The MS-P and MS-W commands SHALL have the same priority. If one of these commands is already issued, and the other command is issued afterwards, it SHALL be ignored. If two LERs are requesting opposite operations simultaneously, i.e. one LER is sending MS-P while the other LER is sending MS-W, the MS-W SHALL be considered to have a higher priority than MS-P, and MS-P SHALL be ignored.

4.3. Updates to Section 3.2. Remote Requests

Replace the following bullet item in the remote request list:

- o Remote MS - indicates that the remote end point is operating under an operator command to switch the traffic from the working path to the protection path.

With:

- o Remote MS-P - indicates that the remote end point is operating under an operator command to switch the traffic from the working path to the protection path.
- o Remote MS-W - indicates that the remote end point is operating under an operator command to switch the traffic from the protection path to the working path.

Replace the following bullet item:

- o Remote DNR - indicates that the remote end point has determined that the failure condition has recovered and will continue transporting traffic on the protection path due to operator configuration that prevents automatic reversion to the Normal state.

With:

- o Remote DNR - indicates that the remote end point has determined that the switch-over condition has ceased or that the failure condition has recovered and will continue transporting traffic on the protection path due to operator configuration that prevents automatic reversion to the Normal state.

4.4. Updates to Section 3.6. PSC Control States

Replace the following bullet item in the protection domain states list:

- o Protecting administrative state - The operator has issued a command switching the user traffic to the protection path.

With:

- o Switching administrative state - The operator has issued a command switching the user traffic either from the working path to the protection path or from the protection path to the working path.

4.5. Updates to Section 4.2.2. PSC Request Field

Replace the following bullet item in the request list:

- o (5) Manual Switch - indicates that the transmitting end point has switched traffic to the protection path as a result of an administrative Manual Switch command. The FPath field SHALL indicate that the working path is being blocked (i.e., FPath set to 1), and the Path field SHALL indicate that user data traffic is being transported on the protection path (i.e., Path set to 1).

With:

- o (5) Manual Switch - indicates that the transmitting end point has switched traffic to the protection path as a result of an administrative Manual Switch to Protection (MS-P) command or to the working path as a result of an administrative Manual Switch to Working (MS-W) command. Two commands, MS-P and MS-W are represented by the same Request Field value, but differentiated by the FPath value. When traffic is switched to the protection path, the FPath field SHALL indicate that the working path is being blocked (i.e., FPath set to 1), and the Path field SHALL indicate that user data traffic is being transported on the protection path (i.e., Path set to 1). When traffic is switched to the working path, the FPath field SHALL indicate that the protection path is being blocked (i.e., FPath set to 0), and the Path field SHALL indicate that user data traffic is being transported on the working path (i.e., Path set to 0).

4.6. Updates to Section 4.3.2. Priority of Inputs

Replace the following number item:

8. Manual Switch (operator command)

With:

8. Manual Switch to Protection/Working (operator command)

Replace the following two paragraphs:

As was noted above, the Local Request logic SHALL always select the local input indicator with the highest priority as the current local request, i.e., only the highest priority local input will be used to affect the control logic. All local inputs with lower priority than this current local request will be ignored.

The remote message from the far-end LER is assigned a priority just below the similar local input. For example, a remote Forced Switch would have a priority just below a local Forced Switch but above a local Signal Fail on protection input. As mentioned in Section 3.6.1, the state transition is determined by the higher priority input between the highest priority local input and the remote message. This also determines the classification of the state as local or remote. The following subsections detail the transition based on the current state and the higher priority of these two inputs.

With:

As was noted above, the Local Request logic SHALL always select the local input indicator with the highest priority as the current local request, i.e., only the highest priority local input will be used to affect the control logic. All local inputs with lower priority than this current local request will be ignored. For local inputs with same priority, first-come, first-served rule is applied. For example, once MS-P (or MS-W) local input is determined as the highest priority local input, then subsequent MS-W (or MS-P) local input will be ignored and automatically canceled.

The remote message from the far-end LER is assigned a priority just below the same local input. For example, a remote Forced Switch would have a priority just below a local Forced Switch but above a local Signal Fail on protection input.

However, if the LER is in a remote state due to a remote message, a subsequent local input having the same priority but requesting different action to the control logic, will be considered as having lower priority than the remote message, and will be ignored. For example, if the LER is in remote Switching administrative status due to a remote MS-P, then subsequent local MS-W will be ignored and automatically canceled.

It should be noted that there is a reverse case where one LER receives a local command and the other LER receives, simultaneously, a command with the same priority but requesting different action. In this case, each of the two LERs receives a subsequent remote message having the same priority but requesting different action, while the LER is in a local state due to the local input. In this case, a priority must be set for the commands with the same priority regardless of its origin (local input or remote message). For example, one LER receives MS-P as a local input and the other LER receives MS-W as a local input, simultaneously. In this case, MS-W SHALL be considered as having higher priority than MS-P at both LERs.

In order to resolve the equal priority conditions described above, following rules are defined:

- a) If two local inputs having same priority but requesting different action come to the Local Request logic, then the input coming first SHALL be considered to have a higher priority than the other coming later (first-come, first-served).
- b) If the LER receives both a local input and a remote message with the same priority and requesting the same action, i.e., the same PSC Request Field and the same FPath value, then the local input SHALL be considered to have a higher priority than the remote message.
- c) If the LER receives both a local input and a remote message with the same priority but requesting different actions, i.e., the same PSC Request Field but different FPath value, then the first-come, first-served rule SHALL be applied. If the remote message comes first, then the state SHALL be a remote state and subsequent local input is ignored. However, if the local input comes first, the first-come, first-served rule cannot be applied and must be viewed as simultaneous condition. This is because the subsequent remote message will not be an acknowledge of the local input by the far-end node. In this case, the priority SHALL be determined by rules for each simultaneous conditions.

- d) If the LER receives both MS-P and MS-W commands either as local input or remote message and the LER is in a local Switching administrative state, then the MS-W command SHALL be considered to have a higher priority than the MS-P command.

As mentioned in Section 3.6.1, the state transition is determined by the higher priority input between the highest priority local input and the remote message. This also determines the classification of the state as local or remote. The following subsections detail the transition based on the current state and the higher priority of these two inputs.

4.7. Updates to Section 4.3.3.1. Normal State

Replace the following bullet item in the reaction to local input list:

- o A local Forced Switch input SHALL cause the LER to go into local Protecting administrative state and begin transmission of an FS(1,1) message.

With:

- o A local Forced Switch input SHALL cause the LER to go into local Switching administrative state and begin transmission of an FS(1,1) message.

Replace the following bullet item in the reaction to local input list:

- o A local Manual Switch input SHALL cause the LER to go into local Protecting administrative state and begin transmission of an MS(1,1) message.

With:

- o A local Manual Switch Protection input SHALL cause the LER to go into local Switching administrative state and begin transmission of an MS(1,1) message.
- o A local Manual Switch Working input SHALL cause the LER to go into local Switching administrative state and begin transmission of an MS(0,0) message.

Replace the following bullet item in the reaction to remote message list:

- o A remote Forced Switch message SHALL cause the LER to go into remote Protecting administrative state and begin transmitting an NR(0,1) message.

With:

- o A remote Forced Switch message SHALL cause the LER to go into remote Switching administrative state and begin transmitting an NR(0,1) message.

Replace the following bullet item in the reaction to remote message list:

- o A remote Manual Switch message SHALL cause the LER to go into remote Protecting administrative state, and transmit an NR(0,1) message.

With:

- o A remote Manual Switch to Protection message SHALL cause the LER to go into remote Switching administrative state, and transmit an NR(0,1) message.
- o A remote Manual Switch to Working message SHALL cause the LER to go into remote Switching administrative state, while continuing to transmit the NR(0,0) message.

4.8. Updates to Section 4.3.3.2. Unavailable State

Replace the following bullet item in the reaction to local input list:

- o A local Forced Switch SHALL be ignored by the PSC Control logic when in Unavailable state as a result of a (local or remote) Lockout of protection. If in Unavailable state due to an SF on protection, then the FS SHALL cause the LER to go into local Protecting administrative state and begin transmitting an FS(1,1) message. It should be noted that due to the unavailability of the protection path (i.e., due to the SF condition) that this FS may not be received by the far-end until the SF condition is cleared.

With:

- o A local Forced Switch SHALL be ignored by the PSC Control logic when in Unavailable state as a result of a (local or remote) Lockout of protection. If in Unavailable state due to an SF on protection, then the FS SHALL cause the LER to go into local Switching administrative state and begin transmitting an FS(1,1)

message. It should be noted that due to the unavailability of the protection path (i.e., due to the SF condition) that this FS may not be received by the far-end until the SF condition is cleared.

Replace the following bullet item in the reaction to remote message list:

- o A remote Forced Switch message SHALL be ignored by the PSC Control logic when in Unavailable state as a result of a (local or remote) Lockout of protection. If in Unavailable state due to a local or remote SF on protection, then the FS SHALL cause the LER to go into remote Protecting administrative state; if in Unavailable state due to local SF, begin transmitting an SF(0,1) message.

With:

- o A remote Forced Switch message SHALL be ignored by the PSC Control logic when in Unavailable state as a result of a (local or remote) Lockout of protection. If in Unavailable state due to a local or remote SF on protection, then the FS SHALL cause the LER to go into remote Switching administrative state; if in Unavailable state due to local SF, begin transmitting an SF(0,1) message.

4.9. Updates to Section 4.3.3.3. Protecting Administrative State

Replace the title of this section with "Switching Administrative State".

Replace the following text in the first paragraph:

In the Protecting administrative state, the user data traffic SHALL be transported on the protection path, while the working path is blocked due to an operator command, i.e., Forced Switch or Manual Switch.

With:

In the Switching administrative state, the user data traffic SHALL be transported on either the protection path or working path, depending on an operator command. If FS or MS-P command is in effect, the working path is blocked and the traffic SHALL be transported on the protection path. If MS-W command is in effect, the protection path is blocked and the traffic SHALL be transported on the working path.

Replace the reaction to local input list with:

- o A local Clear SHALL be ignored if in remote Switching administrative state. If in local Switching administrative state due to local FS or MS-P, then this input SHALL cause the LER to go into Normal state when the LER is configured for revertive behavior, or Do-not-Revert State when the LER is configured for non-revertive behavior. If in local Switching administrative state due to local MS-W, then this input SHALL cause the LER to go into Normal state.
- o A local Lockout of protection input SHALL cause the LER to go into local Unavailable state and begin transmission of an LO(0,0) message.
- o A local Forced Switch input SHALL cause the LER to remain in local Switching administrative state and transmit an FS(1,1) message.
- o A local Signal Fail indication on the protection path SHALL cause the LER to go into local Unavailable state and begin transmission of an SF(0,0) message, if the current state is due to a (local or remote) MS-P or MS-W command. If the LER is in (local or remote) Switching administrative state due to an FS situation, then the SF on protection SHALL be ignored.
- o A local Signal Fail indication on the working path SHALL cause the LER to go into local Protecting failure state and begin transmitting an SF(1,1) message, if the current state is due to a (local or remote) MS-P or MS-W command. If the LER is in remote Switching administrative state due to a remote Forced Switch command, then this local indication SHALL cause the LER to remain in remote Switching administrative state and transmit an SF(1,1) message. If the LER is in local Switching administrative state due to a local Forced Switch command, then this indication SHALL be ignored (i.e., the indication should have been blocked by the Local Request logic).
- o A local Clear SF SHALL clear any local SF condition that may exist. If in remote Switching administrative state, the LER SHALL stop transmitting the SF(x,1) message and begin transmitting an NR(0,1) message.
- o A local Manual Switch to Protection input SHALL be ignored if in remote Switching administrative state due to a remote Forced Switch command. If the current state is due to a (local or remote) Manual Switch to Protection operator command, it SHALL cause the LER to remain in local Switching administrative state and transmit an MS(1,1) message. If the current state is due to a (local or remote) Manual Switch to Working operator command, the local MS-P SHALL be ignored.

- o A local Manual Switch to Working input SHALL be ignored if in remote Switching administrative state due to a remote Forced Switch command. If the current state is due to a (local or remote) Manual Switch to Working operator command, it SHALL cause the LER to remain in local Switching administrative state and transmit an MS(0,0) message. If the current state is due to a (local or remote) Manual Switch to Protection operator command, the local MS-W SHALL be ignored.
- o All other local inputs SHALL be ignored.

Replace the reaction to remote message list with:

- o A remote Lockout of protection message SHALL cause the LER to go into remote Unavailable state and begin transmitting an NR(0,0) message. It should be noted that this automatically cancels the current Forced Switch, Manual Switch to Protection or Manual Switch to Working command and data traffic is reverted to the working path, if required.
- o A remote Forced Switch message SHALL be ignored by the PSC Process logic if there is an active local Forced Switch operator command. If the Switching administrative state is due to a remote Forced Switch message, then the LER SHALL remain in remote Switching administrative state and continue transmitting the last message. If the Switching administrative state is due to either a local or remote Manual Switch to Protection or Manual Switch to Working command, then the LER SHALL remain in remote Switching administrative state (updating the state information with the proper relevant information) and begin transmitting an NR(0,1) message.
- o A remote Signal Fail message indicating a failure on the protection path SHALL cause the LER to go into remote Unavailable state and begin transmitting an NR(0,0) message, if the Switching administrative state is due to a Manual Switch to Protection or Manual Switch to Working command. It should be noted that this automatically cancels the current Manual Switch to Protection or Manual Switch to Working command, and data traffic is reverted to the working path, if required.
- o A remote Signal Fail message indicating a failure on the working path SHALL be ignored if there is an active local Forced Switch command. If the Switching administrative state is due to a local or remote Manual Switch to Protection or Manual Switch to Working, then the LER SHALL go to remote Protecting failure state and begin transmitting an NR(0,1) message.

- o A remote Manual Switch to Protection message SHALL be ignored by the PSC Control logic if in Switching administrative state due to a local or remote Forced Switch. If in Switching administrative state due to a remote Manual Switch to Protection, then the LER SHALL remain in remote Switching administrative state and continue transmitting the current message. If in local Switching administrative state due to an active Manual Switch to Protection, then the LER SHALL remain in local Switching administrative state and continue transmission of the MS(1,1) message. If in Switching administrative state due to a remote MS-W, then the LER SHALL remain in remote Switching administrative state, and begin transmitting an NR(0,1) message. If in Switching administrative state due to a local MS-W, then the remote MS-P message SHALL be ignored.
- o A remote Manual Switch to Working message SHALL be ignored by the PSC Control logic if in Switching administrative state due to a local or remote Forced Switch. If in Switching administrative state due to a remote MS-W, then the LER SHALL remain in remote Switching administrative state and continue transmission of an NR(0,0) message. If in Switching administrative state due to a local MS-W, then the remote MS-W message SHALL be ignored. If in Switching administrative state due to a remote MS-P, then the LER SHALL remain in remote Switching administrative state and begin transmitting an NR(0,0) message. If in Switching administrative state due to a local MS-P, then the LER SHALL go into remote Switching administrative state and begin transmitting an NR(0,0) message. It should be noted that this automatically cancels the current MS-P command.
- o A remote DNR(0,1) message SHALL be ignored if in local Switching administrative state. If in remote Switching administrative state due to a remote FS or MS-P, then the LER SHALL go to Do-not-Revert state and continue transmitting an NR(0,1) message. If in remote Switching administrative state due to a remote MS-W, then the remote DNR message SHALL be ignored.
- o A remote NR(0,0) message SHALL be ignored if in local Switching administrative state. If in remote Switching administrative state due to remote FS and there is no active local Signal Fail indication, then the LER SHALL go into Normal state and begin transmitting an NR(0,0) message. If there is a local Signal Fail on the working path, the LER SHALL go into local Protecting failure state and begin transmitting an SF(1,1) message. If in remote Switching administrative state due to remote MS-P or MS-W, then the LER SHALL go into Normal state and begin transmitting an NR(0,0) message. If in local Switching administrative state due to local MS-P or MS-W, then the remote NR(0,0) message SHALL be

ignored.

- o All other remote messages SHALL be ignored.

4.10. Updates to Section 4.3.3.4. Protecting Failure State

Replace the following bullet item in the reaction to local input list:

- o A local Forced Switch input SHALL cause the LER to go into Protecting administrative state and begin transmission of an FS(1,1) message.

With:

- o A local Forced Switch input SHALL cause the LER to go into Switching administrative state and begin transmission of an FS(1,1) message.

Replace the following bullet item in the reaction to remote message list:

- o A remote Forced Switch message SHALL cause the LER go into remote Protecting administrative state, and if in local Protecting failure state, the LER SHALL transmit the SF(1,1) message; otherwise, it SHALL transmit NR(0,1).

With:

- o A remote Forced Switch message SHALL cause the LER go into remote Switching administrative state, and if in local Protecting failure state, the LER SHALL transmit the SF(1,1) message; otherwise, it SHALL transmit NR(0,1).

4.11. Updates to Section 4.3.3.5. Wait-to-Restore State

Replace the following bullet item in the reaction to local input list:

- o A local Forced Switch command SHALL send the Stop command to the WTR timer, go into local Protecting administrative state, and begin transmission of an FS(1,1) message.

With:

- o A local Forced Switch command SHALL send the Stop command to the WTR timer, go into local Switching administrative state, and begin transmission of an FS(1,1) message.

Replace the following bullet item in the reaction to local input list:

- o A local Manual Switch input SHALL send the Stop command to the WTR timer, go into local Protecting administrative state, and begin transmission of an MS(1,1) message.

With:

- o A local Manual Switch to Protection input SHALL send the Stop command to the WTR timer, go into local Switching administrative state, and begin transmission of an MS(1,1) message.
- o A local Manual Switch to Working input SHALL send the Stop command to the WTR timer, go into local Switching administrative state, and begin transmission of an MS(0,0) message.

Replace the following bullet item in the reaction to remote message list:

- o A remote Forced Switch message SHALL send the Stop command to the WTR timer, go into remote Protecting administrative state, and begin transmission of an NR(0,1) message.

With:

- o A remote Forced Switch message SHALL send the Stop command to the WTR timer, go into remote Switching administrative state, and begin transmission of an NR(0,1) message.

Replace the following bullet item in the reaction to remote message list:

- o A remote Manual Switch message SHALL send the Stop command to the WTR timer, go into remote Protecting administrative state, and begin transmission of an NR(0,1) message.

With:

- o A remote Manual Switch to Protection message SHALL send the Stop command to the WTR timer, go into remote Switching administrative state, and begin transmission of an NR(0,1) message.
- o A remote Manual Switch to Working message SHALL send the Stop command to the WTR timer, go into remote Switching administrative state, and begin transmission of an NR(0,0) message.

4.12. Updates to Section 4.3.3.6. Do-not-Revert State

Replace the first paragraph:

Do-not-Revert state is a continuation of the Protecting failure state when the protection domain is configured for non-revertive behavior. While in Do-not-Revert state, data traffic SHALL continue to be transported on the protection path until the administrator sends a command to revert to Normal state. It should be noted that there is a fundamental difference between this state and Normal -- whereas Forced Switch in Normal state actually causes a switch in the transport path used, in Do-not-Revert state, the Forced Switch just switches the state (to Protecting administrative state) but the traffic would continue to be transported on the protection path! To revert back to Normal state, the administrator SHALL issue a Lockout of protection command followed by a Clear command.

With:

Do-not-Revert state is a continuation of either the Protecting failure state or Switching administrative state due to Forced Switch or Manual Switch to Protection when the protection domain is configured for non-revertive behavior. While in Do-not-Revert state, data traffic SHALL continue to be transported on the protection path until the administrator sends a command to revert to Normal state. When the LER transitions into the Do-not-Revert state, the PSC Control Process SHALL check the persistent state of the local triggers to decide if it should further transition into a new state. If the result of this check is a transition into a new state, the LER SHALL transmit the corresponding message described in this section and SHALL use the data path corresponding to the new state. When the protection domain remains in Do-not-Revert state, the end point SHALL transmit an DNR(0,1) message if the state is local, or an NR(0,1) message if the state is remote, indicating -- Nothing to report and data traffic is being transported on the protection path.

Replace the following bullet item in the reaction to local input list:

- o A local Forced Switch command SHALL cause the LER to go into local Protecting administrative state and begin transmission of an FS(1,1) message.

With:

- o A local Forced Switch command SHALL cause the LER to go into local Switching administrative state and begin transmission of an FS(1,1) message.

Replace the following bullet item in the reaction to local input list:

- o A local Manual Switch input SHALL cause the LER to go into local Protecting administrative state and begin transmission of an MS(1,1) message.

With:

- o A local Manual Switch to Protection input SHALL cause the LER to go into local Switching administrative state and begin transmission of an MS(1,1) message.
- o A local Manual Switch to Working input SHALL cause the LER to go into local Switching administrative state and begin transmission of an MS(0,0) message.

Replace the following bullet item in the reaction to remote message list:

- o A remote Forced Switch message SHALL cause the LER to go into remote Protecting administrative state and begin transmission of an NR(0,1) message.

With:

- o A remote Forced Switch message SHALL cause the LER to go into remote Switching administrative state and begin transmission of an NR(0,1) message.

Replace the following bullet item in the reaction to remote message list:

- o A remote Manual Switch message SHALL cause the LER to go into remote Protecting administrative state and begin transmission of an NR(0,1) message.

With:

- o A remote Manual Switch to Protection message SHALL cause the LER to go into remote Switching administrative state and begin transmission of an NR(0,1) message.

- o A remote Manual Switch to Working message SHALL cause the LER to go into remote Switching administrative state and begin transmission of an NR(0,0) message.

4.13. Updates to Appendix A. PSC State Machine Tables

Modify the state machine as follows (only modified cells are shown):

Part 1: Local input state machine

	OC	LO	SF-P	FS	SF-W	SF _c
N				SA:F:L		
UA:LO:L				SA:F:L		
UA:P:L				SA:F:L		
UA:LO:R				SA:F:L		
UA:P:R				SA:F:L		
PF:W:L				SA:F:L		
PF:W:R				SA:F:L		
SA:F:L	[20]					
SA:MW:L	N	UA:LO:L	UA:P:L	SA:F:L	PF:W:L	i
SA:MP:L	[20]			SA:F:L		
SA:F:R				SA:F:L		
SA:MW:R	i	UA:LO:L	UA:P:L	SA:F:L	PF:W:L	i
SA:MP:R				SA:F:L		
WTR				SA:F:L		
DNR				SA:F:L		

	MS-W	MS-P	WTRExp
N	SA:MW:L	SA:MP:L	
UA:LO:L	i		
UA:P:L	i		
UA:LO:R	i		
UA:P:R	i		
PF:W:L	i		
PF:W:R	i		
SA:F:L	i		
SA:MW:L	i	i	i
SA:MP:L	i		
SA:F:R	i		
SA:MW:R	SA:MW:L	i	i
SA:MP:R	i	SA:MP:L	
WTR	i	SA:MP:L	
DNR	SA:MW:L	SA:MP:L	

Part 2: Remote messages state machine

	LO	SF-P	FS	SF-W	MS-W	MS-P
N			SA:F:R		SA:MW:R	SA:MP:R
UA:LO:L					i	
UA:P:L					i	
UA:LO:R					i	
UA:P:R			SA:F:R		i	
PF:W:L			SA:F:R		i	
PF:W:R			SA:F:R		i	
SA:F:L					i	
SA:MW:L	UA:LO:R	UA:P:R	SA:F:R	[13]	i	i
SA:MP:L			SA:F:R		SA:MW:R	
SA:F:R					i	
SA:MW:R	UA:LO:R	UA:P:R	SA:F:R	[13]	i	SA:MP:R
SA:MP:R			SA:F:R		SA:MW:R	
WTR			SA:F:R		SA:MW:R	SA:MP:R
DNR			SA:F:R		SA:MW:R	SA:MP:R

	WTR	DNR	NR
N			
UA:LO:L			
UA:P:L			
UA:LO:R			
UA:P:R			
PF:W:L			
PF:W:R			
SA:F:L			
SA:MW:L	i	i	i
SA:MP:L			
SA:F:R			
SA:MW:R	i	i	i
SA:MP:R			
WTR			
DNR			

Replace the following item in the footnotes for the table:

[4] Remain in the current state (PA:F:R) and transmit SF(1,1).

[8] Remain in PA:F:R and transmit NR(0,1).

[19] Transition to PA:F:R and send SF (0,1).

With:

[4] Remain in the current state (SA:F:R) and transmit SF(1,1).

[8] Remain in SA:F:R and transmit NR(0,1).

[19] Transition to SA:F:R and send SF(0,1).

Add the following item in the footnotes for the table:

[20] If domain configured for revertive behavior transition to N,
else transition to DNR.

5. Security considerations

No specific security issue is raised in addition to those ones
already documented in [RFC6378]

6. IANA considerations

This document makes no request of IANA.

Note to RFC Editor: this section may be removed on publication as an RFC.

7. Acknowledgements

8. References

8.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC4427] Mannie, E. and D. Papadimitriou, "Recovery (Protection and Restoration) Terminology for Generalized Multi-Protocol Label Switching (GMPLS)", RFC 4427, March 2006.
- [RFC5654] Niven-Jenkins, B., Brungard, D., Betts, M., Sprecher, N., and S. Ueno, "Requirements of an MPLS Transport Profile", RFC 5654, September 2009.
- [RFC6378] Weingarten, Y., Bryant, S., Osborne, E., Sprecher, N., and A. Fulignoli, "MPLS Transport Profile (MPLS-TP) Linear Protection", RFC 6378, October 2011.

8.2. Informative References

- [LIAISON1205] ITU-T SG15, "Liaison Statement: Recommendation ITU-T G.8131/Y.1382 revision - Linear protection switching for MPLS-TP networks", <https://datatracker.ietf.org/liaison/1205/> , October 2012.
- [LIAISON1234] ITU-T SG15, "Liaison Statement: Recommendation ITU-T G.8131 revision - Linear protection switching for MPLS-TP networks", <https://datatracker.ietf.org/liaison/1234/> , February 2013.

Authors' Addresses

Taesik Cheung
ETRI
218 Gajeongno
Yuseong-gu, Daejeon 305-700
South Korea

Phone: +82-42-860-5646
Email: cts@etri.re.kr

Alessandro D'Alessandro
Telecom Italia
via Reiss Romoli, 274
Torino 10141
Italy

Phone: +39 011 2285887
Email: alessandro.dalessandro@telecomitalia.it

Huub van Helvoort
Huawei Technologies
Karspeldreef 4
Amsterdam 1101 CJ
The Netherlands

Phone: +31 20 4300832
Email: huub.van.helvoort@huawei.com

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: January 16, 2014

M. Chen
X. Xu
Z. Li
Huawei
L. Fang
Cisco
July 15, 2013

MultiProtocol Label Switching (MPLS) Source Label
draft-chen-mpls-source-label-00

Abstract

An MultiProtocol Label Switching (MPLS) label is originally defined to identify a Forwarding Equivalence Class (FEC), a packet is assigned to a specific FEC based on its network layer destination address. It's difficult or even impossible to derive the source information from the label. For some applications, source identification is a critical requirement. For example, performance monitoring, traffic matrix measurement and collection, where the monitoring node needs to identify where a packet was sent from.

This document introduces the concept of Source Label (SL) that is carried in the label stack and used to identify the ingress Label Switching Router (LSR) of an Label Switched Path (LSP).

Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [RFC2119].

Status of This Memo

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

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

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

This Internet-Draft will expire on January 16, 2014.

Copyright Notice

Copyright (c) 2013 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. Problem Statement and Introduction	3
2. Source Label	4
3. Use Cases	4
3.1. Performance Measurement	4
3.2. Traffic Matrix Measurement and Steering	5
4. Data Plane Processing	6
4.1. Ingress LSR	6
4.2. Transit LSR	6
4.3. Egress LSR	7
4.4. Penultimate Hop LSR	7
5. Source Label Signaling	7
5.1. Source Label Capability Signaling	7
5.1.1. LDP Extensions	7
5.1.2. BGP Extensions	8
5.1.3. RSVP-TE Extensions	9
5.2. Source Label Distribution	9
6. IANA Considerations	9
6.1. Source Label Indication	9
6.2. LDP Source Label Capability TLV	9
6.3. BGP Source Label Capability Attribute	10
6.4. RSVP-TE Source Label Capability	10
7. Security Considerations	10
8. Acknowledgements	10
9. References	10
9.1. Normative References	10
9.2. Informative References	11
Authors' Addresses	11

1. Problem Statement and Introduction

An MultiProtocol Label Switching (MPLS) label [RFC3031] is originally defined for packet forwarding and assumes the forwarding/destination address semantics. As no source address information is carried in the label stack, there is no way to directly derive the source address information from the label or label stack.

MPLS LSPs can be categorized into four different types:

Point-to-Point (P2P)

Point-to-Multipoint (P2MP)

Multipoint-to-Point (MP2P)

Multipoint-to-Multipoint (MP2MP)

LSPs that are established by the Resource Reservation Protocol Traffic Engineering (RSVP-TE) [RFC3209] and Pseudowires (PWs) belong to P2P or P2MP types. LSPs that are established by the classic Label Distribution Protocol (LDP) [RFC5036], Layer 3 Private Network (L3VPN) and Virtual Local Area Network (VPLS) LSPs belong to MP2P or MP2MP types.

For those LSPs belong to the MP2P and MP2MP types, it is not possible to derive the source address information from the label. For the P2P or P2MP LSPs, the source address information may be implicitly derived from the label (e.g., P2P or P2MP LSPs established by RSVP-TE), but it requires that some further information is used (e.g., control plane information). However, this is not always possible for all P2P LSPs. One example is the Multi-Segment Pseudowire (MS-PW), it is impossible to derive the source address information from the PW label. Because an MS-PW label assumes the forwarding and destination address semantics which is quite different from the source address semantics that a Single-Segment Pseudowire (SS-PW) label assumes.

Comparing to the pure IP forwarding where both source and destination addresses are encoded in the IP packet header, the essential issue of the MPLS encoding is that the label stack does not explicitly include any source address information, i.e., a Source Label (SL). For some applications, source identification is a critical requirement. For example, performance monitoring, the monitoring nodes need to identify where packets were sent from and then can count the packets according to some constraints. In addition, traffic matrix measurement and collection is the precondition of traffic steering, and capable of traffic steering is an important requirement of Software Defined Network (SDN). To measure and collect traffic matrix information, the source address information is necessary.

This document introduces the concept of a Source Label. A SL uniquely identifies a node within an administrative domain, it is carried in the label stack and used to identify the ingress LSR(s) of an LSP.

2. Source Label

A Source Label is defined to uniquely identify a node that is (one of) the ingress LSR(s) to a specific LSP. In its function as a Source Label (ingress node identifier), it MUST be unique within a domain. In cases where a Source Label is used across domains it MUST be unique within the scope it is used. How to guarantee the uniqueness of Source Labels is out of scope for this document. Source Labels are not used for forwarding.

In order to indicate whether a label is a source label, a Source Label Indicator (SLI) is introduced. The SLI is a reserved label that is placed immediately before the source label in the label stack, which is used to indicate that the next label in the label stack is a source label. The value of SLI is TBD1.

3. Use Cases

3.1. Performance Measurement

There are two typical types of performance measurement: one is active performance measurement, and the other is passive performance measurement.

In active performance measurement the receiver measures the injected packets to evaluate the performance of a path. The active measurement measures the performance of the extra injected packets. The IP Performance Metrics (IPPM) working group has defined specifications for the active performance measurement.

In passive performance measurement, no artificial traffic is injected into the flow and measurements are taken to record the performance metrics of the real traffic. The Multiprotocol Label Switching (MPLS) PM protocol [RFC6374] for packet loss is an example of passive performance measurement. For a specific receiver, in order to count the received packets of a flow, it has to know whether a received packet belongs to which target flow under test and the source identification is a critical condition.

As discussed in the previous section, the existing MPLS label or label stack do not carry the source information. So, for an LSP, the ingress LSR can put a source label in the label stack, and then the egress LSR can use the source label for packets identifying and counting.

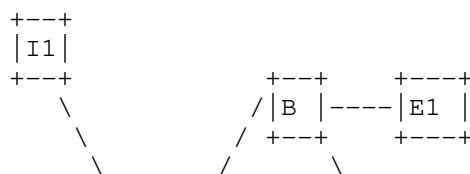
3.2. Traffic Matrix Measurement and Steering

A Traffic Matrix (TM) provides, for every ingress node (i) into the network and every egress node (j) out of the network, the volume of traffic $T(i,j)$ from i to j over a given time interval.

Since the ingress node knows the source and destination of the traffic, it's normal to measure the traffic matrix at every ingress node. But in some scenarios, it may need to measure the traffic at the egress or intermediate nodes. Taking Figure 1 as an example, from the west to east point of view, there are three ingress nodes (I1, I2 and I3) and three egress nodes (E1, E2 and E3), A, B and C are intermediate nodes. It is not necessary to measure the traffic matrix of the whole network all the time, it sometimes just wants to know the received traffic matrix of a specific egress node (e.g., E2). So, to measure received traffic matrix at node E2 would be then a better choice.

In addition, for an intermediate node (e.g., A), it may need to measure the transmitted traffic hence to steer some traffic from the congestion path to idle path.

Wherever at egress or intermediate node, source identification is necessary. The ingress LSR can put the source label into the label stack to help the egress and intermediate LSR to identify and measure the traffic.



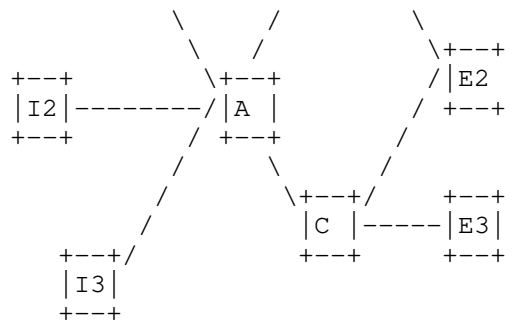


Figure 1: Traffic Matrix Measurement and Steering

4. Data Plane Processing

4.1. Ingress LSR

For an LSP, the ingress LSR MUST make sure that the egress LSR is able to process the Source Label before inserting a SL and SLI into the label stack. Therefore, an egress LSR SHOULD signal (see Section 5.1) to the ingress LSR whether it is able to process the Source Label. Once the ingress LSR knows that the egress LSR can process Source Label, it can choose whether or not to insert the SL and SLI into the label stack.

When a SL to be included in a label stack, the steps are as follows:

1. Push the SL label, the Bos bit for the SL depends on whether the SL is the bottom label;
2. Push the SLI, the TTL and TC field for the SLI SHOULD be set to the same values as for the LSP Label (L);
3. Push the LSP Label (L) .

Then the label stack looks like: <...L, SLI, SL...>. There may be multiple pairs of SLI and SL inserted into the label stack, each pair is related to an LSP. For an LSP, only one pair of SLI and SL SHOULD be inserted.

4.2. Transit LSR

There is no change in forwarding behavior for transit LSRs. But if a transit LSR can recognize the SLI, it may use the SL to collect traffic throughput and/or measure the performance of the LSP.

4.3. Egress LSR

When an egress LSR receives a packet with a SLI/SL pair, if the egress LSR is able to process the SL; it pops the LSP label (if have), SLI and SL; then processes remaining packet header as normal. If the egress LSR is not able to process the SL, the packet SHOULD be dropped.

4.4. Penultimate Hop LSR

The penultimate hop LSR MUST not pop the SLI and SL.

5. Source Label Signaling

Source label signaling includes two aspects: one is source label capability signaling, the other is source label distribution.

5.1. Source Label Capability Signaling

Before inserting a source label in the label stack, an ingress LSR MUST know whether the egress LSR is able to process the source label. Therefore, an egress LSR should signal to the ingress LSRs its ability to process the Source Label. This is called Source Label Capability (SLC), it is very similar to the "Entropy Label Capability (ELC)" [RFC6790].

5.1.1. LDP Extensions

A new LDP TLV [RFC5036], SLC TLV, is defined to signal an egress's ability to process source label. The SLC TLV may appear as an Optional Parameter of the Label Mapping Message. The presence of the SLC TLV in a Label Mapping Message indicates to ingress LSRs that the egress LSR can process source labels for the associated LSP.

The structure of the SLC TLV is shown below.

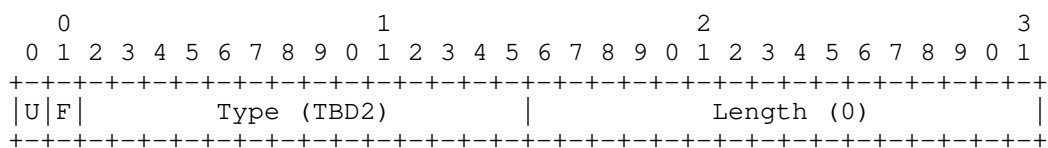


Figure 1: Source Label Capability TLV

This U bit MUST be set to 1. If the SLC TLV is not understood by the receiver, then it MUST be ignored.

This F bit MUST be set to 1. Since the SLC TLV is going to be propagated hop-by-hop, it should be forwarded even by nodes that may not understand it.

Type: TBD2.

Length field: This field specifies the total length in octets of the SLC TLV and is defined to be 0.

An LSR that receives a Label Mapping with the SLC TLV but does not understand it MUST propagate it intact to its neighbors and MUST NOT send a notification to the sender (following the meaning of the U- and F-bits). An LSR X may receive multiple Label Mappings for a given FEC F from its neighbors. In its turn, X may advertise a Label Mapping for F to its neighbors. If X understands the SLC TLV, and if any of the advertisements it received for FEC F does not include the SLC TLV, X MUST NOT include the SLC TLV in its own advertisements of F. If all the advertised Mappings for F include the SLC TLV, then X MUST advertise its Mapping for F with the SLC TLV. If any of X's neighbors resends its Mapping, sends a new Mapping or sends a Label Withdraw for a previously advertised Mapping for F, X MUST re-evaluate the status of SLC for FEC F, and, if there is a change, X MUST re-advertise its Mapping for F with the updated status of SLC.

5.1.2. BGP Extensions

When Border Gateway Protocol (BGP) [RFC4271] is used for distributing Network Layer Reachability Information (NLRI) as described in, for example, [RFC3107], [RFC4364], the BGP UPDATE message may include the SLC attribute as part of the Path Attributes. This is an optional, transitive BGP attribute of value TBD3. The inclusion of this attribute with an NLRI indicates that the advertising BGP router can process source labels as an egress LSR for all routes in that NLRI.

A BGP speaker S that originates an UPDATE should include the SLC attribute only if both of the following are true:

A1: S sets the BGP NEXT_HOP attribute to itself AND

A2: S can process source labels.

Suppose a BGP speaker T receives an UPDATE U with the SLC attribute. T has two choices. T can simply re-advertise U with the SLC attribute if either of the following is true:

B1: T does not change the NEXT_HOP attribute OR

B2: T simply swaps labels without popping the entire label stack and processing the payload below.

An example of the use of B1 is Route Reflectors. However, if T changes the NEXT_HOP attribute for U and in the data plane pops the entire label stack to process the payload, T MAY include an SLC attribute for UPDATE U' if both of the following are true:

C1: T sets the NEXT_HOP attribute of U' to itself AND

C2: T can process source labels. Otherwise, T MUST remove the SLC attribute.

5.1.3. RSVP-TE Extensions

Source label support is signaled in RSVP-TE [RFC3209] using the Source Label Capability (SLC) flag in the Attribute Flags TLV of the LSP_ATTRIBUTES object [RFC5420]. The presence of the SLC flag in a Path message indicates that the ingress can process entropy labels in the upstream direction; this only makes sense for a bidirectional LSP and MUST be ignored otherwise. The presence of the SLC flag in a Resv message indicates that the egress can process entropy labels in the downstream direction. The bit number for the SLC flag is TBD4.

5.2. Source Label Distribution

Based on the Source Label, an egress or intermediate LSR can identify from where an MPLS packet is sent. To achieve this, the egress and/or intermediate LSRs have to know which ingress LSR is related to which Source Label before using the Source Label to derive the source information. Therefore, there needs a mechanism to distribute the mapping information between an ingress LSR and its Source Label. For example, defines extensions to LDP, BGP, RSVP-TE and/or Interior Gateway Protocol (IGP) to distribute to source label. The source label distribution will be defined in future revision or another document.

6. IANA Considerations

6.1. Source Label Indication

IANA is required to allocate a reserved label (TBD1) for the Source Label Indicator (SLI) from the "Multiprotocol Label Switching Architecture (MPLS) Label Values" Registry.

6.2. LDP Source Label Capability TLV

IANA is required to allocate a value of TBD2 from the IETF Consensus range (0x0001-0x07FF) in the "TLV Type Name Space" registry as the "Source Label Capability TLV".

6.3. BGP Source Label Capability Attribute

IANA is required to allocate a Path Attribute Type Code TBD3 from the "BGP Path Attributes" registry as the "BGP Source Label Capability Attribute".

6.4. RSVP-TE Source Label Capability

IANA is required to allocate a new bit from the "Attribute Flags" sub-registry of the "Resource Reservation Protocol-Traffic Engineering (RSVP-TE) Parameters" registry.

Bit	Name	Attribute	Attribute	RRO
No		Flags Path	Flags Resv	
TBD4	Source Label Capability	Yes	Yes	No

7. Security Considerations

TBD.

8. Acknowledgements

The process of "Source Label Capability Signaling" is largely referred to the process of "ELC signaling"[RFC6790].

The authors would like to thank Carlos Pignataro, Loa Andersson 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.
- [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.
- [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.
- [RFC5420] Farrel, A., Papadimitriou, D., Vasseur, JP., and A. Ayyangarps, "Encoding of Attributes for MPLS LSP Establishment Using Resource Reservation Protocol Traffic Engineering (RSVP-TE)", RFC 5420, February 2009.
- [RFC6374] Frost, D. and S. Bryant, "Packet Loss and Delay Measurement for MPLS Networks", RFC 6374, September 2011.

9.2. Informative References

- [RFC4271] Rekhter, Y., Li, T., and S. Hares, "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, January 2006.
- [RFC6790] Kompella, K., Drake, J., Amante, S., Henderickx, W., and L. Yong, "The Use of Entropy Labels in MPLS Forwarding", RFC 6790, November 2012.

Authors' Addresses

Mach(Guoyi) Chen
Huawei

Email: mach.chen@huawei.com

Xiaohu Xu
Huawei

Email: xuxiaohu@huawei.com

Zhenbin Li
Huawei

Email: lizhenbin@huawei.com

Internet-Draft

Source Label

July 2013

Luyuan Fang
Cisco

Email: lufang@cisco.com

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: January 09, 2014

WQ. Cheng
L. Wang
H. Li
China Mobile
K. Liu
J. He
Huawei Technologies Co., Ltd.
F. Li
China Academy of Telecommunication Research, MIIT., China
J. Yang
ZTE Corporation P.R.China
JF. Wang
Fiberhome Telecommunication Technologies Co., LTD.
July 08, 2013

MPLS-TP Shared-Ring protection (MSRP) mechanism for ring topology
draft-cheng-mpls-tp-shared-ring-protection-01

Abstract

This document describes requirements and solutions for MPLS-TP Shared Ring Protection (MSRP) in the ring topology for point-to-point (P2P) services. The mechanism of MSRP is illustrated and analyzed how it satisfies the requirements in RFC5654 for optimized ring protection.

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 January 09, 2014.

Copyright Notice

Copyright (c) 2013 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	2
1.1. Requirements for MPLS-TP ring protection	3
1.1.1. Recovery for Multiple failures	3
1.1.2. Smooth Upgrade from linear protection to ring protection	4
1.1.3. Configuration complexity	4
1.2. Terminology and Notation	5
1.3. Contributing Authors	5
2. Shared-ring protection for P2P	5
2.1. Basic concept	5
2.1.1. The establishment of the Ring tunnels	6
2.1.2. The distribution and management of ring labels	7
2.1.3. Failure detection	8
2.2. P2P wrapping	9
2.2.1. Wrapping for Link Failure	9
2.2.2. Wrapping for node Failure	10
2.3. P2P short wrapping	10
2.4. P2P steering	12
2.5. P2P wrapping for Interconnected Rings	14
2.5.1. Interconnected ring topology	14
2.5.2. Interconnected ring protection scheme	15
3. Coordination protocol	20
4. Conclusions	20
5. IANA Considerations	20
6. Security Considerations	20
7. References	20
Authors' Addresses	21

1. Introduction

As described in 2.5.6.1. ring protection of MPLS-TP requirements [RFC5654], several service providers have expressed much interest in operating MPLS-TP in ring topologies and required a high-level survivability function in these topologies. In operation network deployment, MPLS-TP networks are often constructed with ring topologies. It calls for an efficient and optimized ring protection mechanism to achieve simplified operation and fast recovery performance.

The requirements for MPLS-TP [RFC5654] state that recovery mechanisms which are optimized for ring topologies could be further developed if it can provide the following features:

- a. Minimize the number of OAM entities for protection
- b. Minimize the number of elements of recovery
- c. Minimize the required label number
- d. Minimize the amount of control and management-plane transactions
- e. Minimize the impact on information exchange if the control plane supports

This document specifies MPLS-TP Shared-Ring Protection mechanisms which can meet all those requirements on ring protection listed in [RFC5654].

This document focus on the solutions for point-to-point transport path. The solution for point-to-multipoint transport is under study and will be presented in a separate document. The basic concept stated in this document also apply to point-multipoint transport path.

1.1. Requirements for MPLS-TP ring protection

The requirements for MPLS-TP ring protection are specified in RFC5654. This document elaborates the requirements in detail.

1.1.1. Recovery for Multiple failures

MPLS-TP is expected to be used in carrier grade metro networks and backbone networks to provide mobile backhaul, carry business customers' services and etc., in which the network survivability is very important. According to R106 B in RFC5654, MPLS-TP recovery mechanisms in a ring SHOULD protect against multiple failures. The following context provides some more detailed illustration about "multiple failures". In metro and backbone networks, the single risk

factor often affects multiple links or nodes. Some examples of risk factors are given as follows:

- multiple links using fibers in one cable or pipeline
- Several nodes shared one power supply system
- weather sensitive micro-wave system

Once one of the above risk factors happens, multiple links or nodes failures may occur simultaneously and those failed links or nodes may locate on a single ring as well as on interconnected rings. Ring protection against multiple failures should cover both multiple failures on a single ring and multiple failures on interconnected rings.

1.1.2. Smooth Upgrade from linear protection to ring protection

It is beneficial for service providers to upgrade protection scheme from linear protection to ring protection in their MPLS-TP network without service interruption. In-service insertion and removal of a node on the ring should also be supported. Therefore, the MPLS-TP ring protection mechanism is supposed to be developed and optimized to comply with this smooth upgrading principle.

1.1.3. Configuration complexity

While deploying linear protection in MPLS-TP networks, the configuration effort of protection depends on the quantity of the services carried. In some large metro networks with more than ten thousand services access, the LSP linear protection capabilities of the metro core nodes should be large enough to meet the network planning requirements, which also leads to the complexity of network protection configuration and operation. While ring protection can reduce the dependency of configuration on the quantity of services, it will simplify the network protection configuration and operation effort. In the application scenarios of deploying linear protection in MPLS-TP network, the configuration of protection has close relationship with the services, LSP quantities. Especially in some large metro networks with more than ten thousands of services access node, the LSP linear protection capabilities of the metro core nodes should be large enough to meet the network planning requirements, which also leads to the complexity of network protection configurations and operations. While the ring protection is based on the mechanisms on section layer, it has loose relationship with the services quantities which could simplify the network protection configurations and operations effort.

1.2. Terminology and Notation

The following syntax will be used to describe the contents of 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.

It should be noted that the label stack may contain additional layers. However, we only present the layers that are related to the protection mechanism.

3. If the Label is assigned by Node x, the Node Name will enclosed in bracket ("()")

1.3. Contributing Authors

Wen Ye (China Mobile)

2. Shared-ring protection for P2P

2.1. Basic concept

This document introduces a novel logic layer of the ring for both working path and protection path for shared ring protection in MPLS-TP networks. As shown in Figure 1, the new logic layer is a ring tunnel on top of the working path or the protection path, namely working ring tunnel and protection ring tunnel respectively. Once a ring tunnel is established, the configuration, management and protection of the ring are all based on the ring tunnel. One port can carry multiple ring tunnels, while one ring tunnel can carry multiple LSPs.

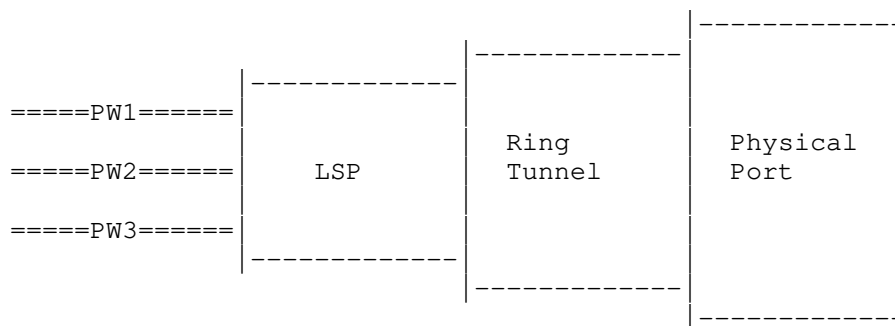


Figure 1 the logic layers of the ring

The label stack used in MPLS-TP Shared Ring Protection mechanism is shown as below.

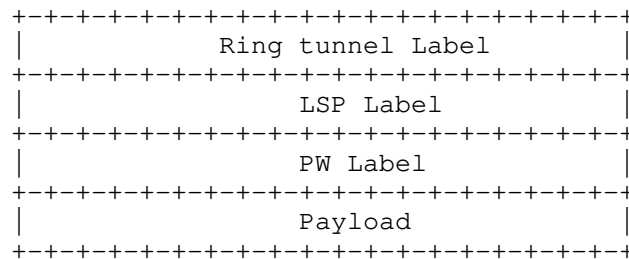


Figure 2 Label stack used in MPLS-TP Shared Ring Protection

2.1.1. The establishment of the Ring tunnels

LSPs which have same exit node share the same ring tunnel. The exit node is the node where the traffic leaves the ring. In other words, all the LSPs that traverse the ring and exit from the same node share the same working ring tunnel and protection ring tunnel. For each exit node, four ring tunnels are established:

- one clockwise working ring tunnel, which is protected by the following protection tunnel,
- one anticlockwise protection ring tunnel,
- one anticlockwise working ring tunnel, which is protected by the following protection tunnel,
- one clockwise protection ring tunnel.

An example is shown in Figure 3 where Node D is the exit node. LSP 1, LSP 2 and LSP 3 enter the ring from Node E, Node A and Node B, respectively, and all leave the ring from Node D. To protect these LSPs that traverse the ring, a clockwise working ring tunnel (RcW_D) via E->F->A->B->C->D, and its protection ring tunnel in the reverse direction (RaP_D) via D->C->B->A->F->E->D are established, respectively; Also, an anti-clockwise working ring tunnel (RaW_D) via C->B->A->F->E->D, and its clockwise protection ring tunnel (RcP_D) via D->E->F->A->B->C->D are established, respectively. Figure 3 only shows RcW_D and RaP_D. A similar provisioning should be applied for any other node on the ring. For other nodes in Figure 3 when acting as an exit node, the ring tunnels are created as follows:

To Node A: RcW_A, RaW_A, RcP_A, RaP_A;

To Node B: RcW_B, RaW_B, RcP_B, RaP_B;

To Node C: RcW_C, RaW_C, RcP_C, RaP_C;

To Node E: RcW_E, RaW_E, RcP_E, RaP_E;

To Node F: RcW_F, RaW_F, RcP_F, RaP_F;

For exit Node D, two working ring tunnels, RcW_D and RaW_D, are terminated on Node D, and two protection ring tunnels, RcP_D and RaP_D, are started from Node D. That means through these working ring tunnels with protection ring tunnels, LSPs which enter the ring from Node D can reach any other nodes on the ring, while Node D can also receive the traffic from any other nodes.

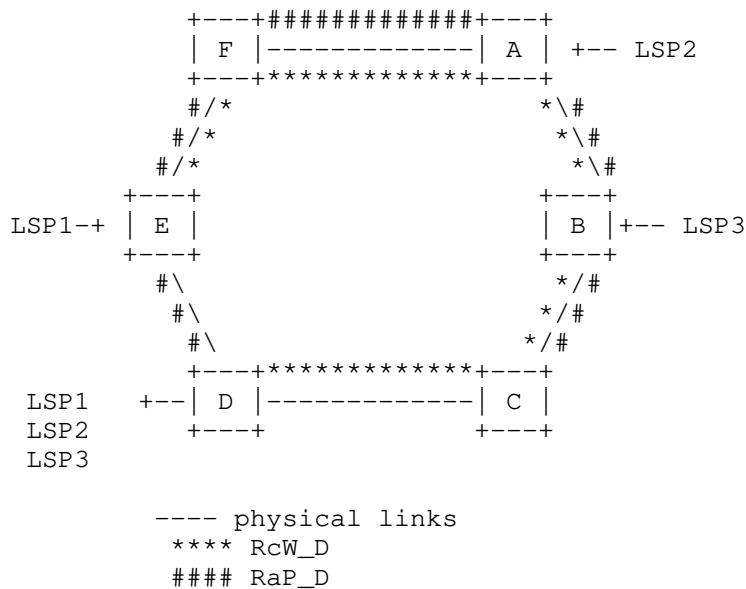


Figure 3 Ring tunnels in MSRP

2.1.2. The distribution and management of ring labels

Ring tunnel labels are distributed by means of downstream-assigned mechanism as defined in [RFC3031]. When a MPLS-TP transport path, such as LSP, enters the ring, the ingress node pushes the working ring tunnel label and sends the traffic to the next hop according to the ring ID and the exit node. The transit nodes within the working ring tunnel swap ring tunnel labels and forward the packets to the next hop; When arriving at the egress node, the egress node removes

the ring tunnel label and forwards the packets based on the inner LSP label and PW label. Figure 4 shows the label operation in the MPLS-TP shared ring protection mechanism. Assume that LSP 1 enters the ring at Node A and exits from Node D, and the following label operations are executed.

1. The traffic LSP1 arrives at Node A with a label stack [LSP1] and is supposed to be forwarded in the clockwise direction of the ring. The clockwise working ring tunnel label RcW_D will be pushed at Node A, the label stack for the forwarded packet at Node A is changed to [RcW_D(B) | LSP1]
2. Transit nodes, in this case, Node B and Node C forward the packets by swapping the working ring tunnel labels. For example, the label [RcW_D(B) | LSP1] is swapped to [RcW_D(C) | LSP1] at Node B.
3. When the packet arrives at Node D (i.e. egress node) with label stack [RcW_D(D) | LSP1], Node D removes RcW_D(D), and subsequently deals with the inner labels of LSP1.
4. All the LSPs which exit from the same node share the same set of ring tunnel labels.

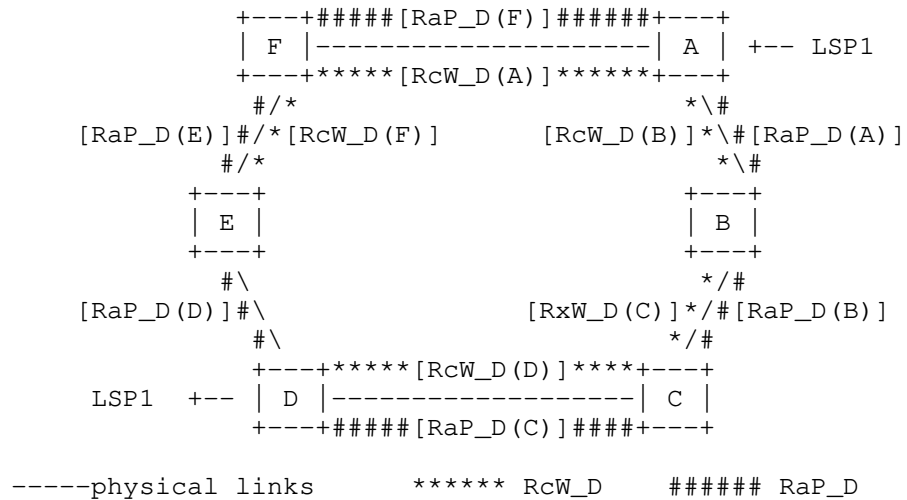


Figure 4 Label operation of MSRP

2.1.3. Failure detection

The MPLS-TP section layer OAM is used to monitor the connectivity between each two adjacent nodes on the ring using the mechanisms

defined in [RFC6371]. Protection switching is triggered by the failure detection in a link in the ring monitored by OAM functions.

Two end ports of a link form an MEG, and an MEG end point (MEP) function is installed in each ring port. CC-V OAM packets are periodically exchanged between each pair of MEPs to monitor the link health. Consecutive losses of CC-V packets (3 packets) will be interpreted as a link failure.

A node failure is regarded as the failure of two links attached to the node. The two nodes adjacent to the failed node detect the failure in the links that are connected to the failed node.

2.2. P2P wrapping

Normal state is shown in Figure 4. The clockwise LSP1 towards node D enters the ring at Node A. In normal state, LSP 1 follows the path A->B->C->D, label operation is [LSP1](original data traffic carried by LSP 1)->[RCW_D(B) | LSP1] (NodeA)->[RCW_D(C) | LSP1] (NodeB)->[RCW_D(D) | LSP1] (NodeC)->[LSP1] (data traffic carried by LSP 1). Then traffic packet will be forwarded based on LSP1 at nodeD.

2.2.1. Wrapping for Link Failure

When a link failure between Node B and Node C occurs, both Node B and Node C detect the failure by OAM mechanism. Node B switches the clockwise working ring tunnel (RcW_D) to the anticlockwise protection ring tunnel (RaP_D) and Node C switches anticlockwise protection ring tunnel (RaP_D) to the clockwise work ring tunnel (RcW_D). The data traffic which enters the ring at Node A and exits at Node D follows the path A->B->A->F->E->D->C->D. The label operation is [LSP1] (Original data traffic)-> [RcW_D(B) | LSP1] (Node A)-> [RaP_D(A) | LSP1] (Node B)->[RaP_D(F) | LSP1] (Node A)->[RaP_D(E) | LSP1] (Node F)->[RaP_D(D) | LSP1] (Node E)-> [RaP_D(C) | LSP1] (Node D)-> [RcW_D(D) | LSP1] (Node C)->[LSP1] (Data traffic exits the ring).

```

+---+##### [RaP_D (F) ] #####+---+
| F |-----| A | +--- LSP1
+---+***** [RcW_D (A) ] *****+---+
#/*                                     *\#
[RaP_D (E) ] #/* [RcW_D (F) ]          [RcW_D (B) ] *\# RaP_D (A)
#/*                                     *\#
+---+                                     +---+
| E |                                     | B |
+---+                                     +---+
#\\                                     *x#
[RaP_D (D) ] #\\                          [RcW_D (C) ] *x# RaP_D (B)
#\\                                     *x#

```

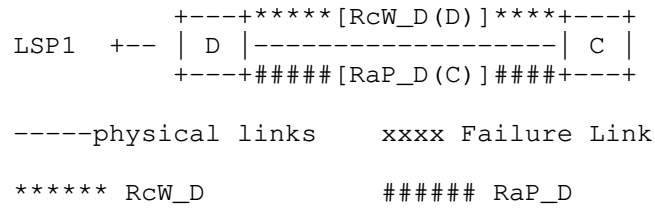


Figure 5 P2P wrapping for link failure in a single ring

2.2.2. Wrapping for node Failure

When Node B fails, Node A detects the failure between A and B and switches the clockwise work ring tunnel(RcW_D) to the anticlockwise protection ring tunnel(RaP_D), Node C detects the failure between C and B and switches the anticlockwise protection ring tunnel(RaP_D) to the clockwise working ring tunnel(RcW_D). The data traffic which enters the ring at Node A and exits at Node D follows the path A->F->E->D->C->D. The label operation is [LSP1](original data traffic carried by LSP 1)-> [RaP_D(F) | LSP1] (NodeA)->[RaP_D(E) | LSP1] (NodeF)-> [RaP_D(D) | LSP1] (NodeE)-> [RaP_D(C) | LSP1] (NodeD)->[RcW_D(D) | LSP1] (NodeC)->[LSP1] (data traffic carried by LSP 1).

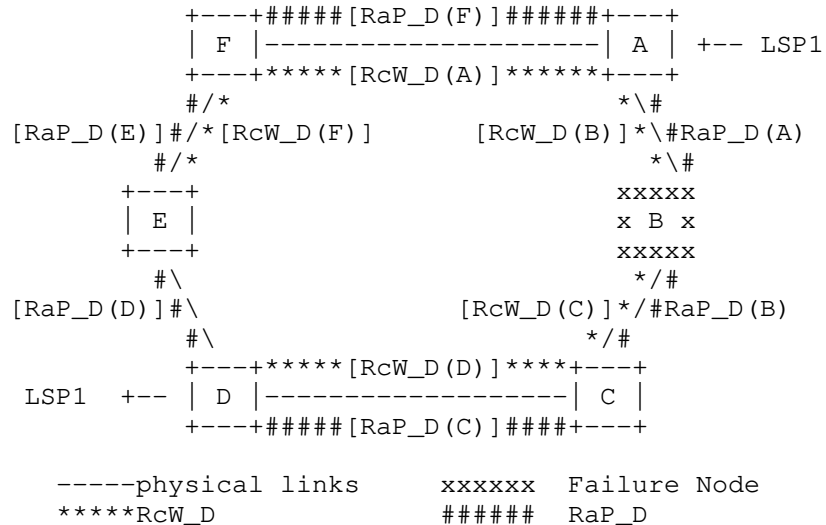


Figure 6 P2P wrapping for node failure in a single ring

2.3. P2P short wrapping

For traditional wrapping protection scheme, Protection switching execute at both nodes neighbored failure respectively , so the traffic will be wrapped twice. This mechanism will cause more latency and bandwidth consume when traffic switched to protection path.

For Short wrapping protection, switching only execute at up-stream node neighbored failure node, and exited ring in protection ring tunnel. This scheme can optimized latency and bandwidth consume when traffic switched to protection path.

In traditional wrapping solution, protection ring tunnel is a closed path in normal state, while in short wrapping solution, protection ring tunnel will remove at exit node. Short wrapping is easy to implement in shared ring protection because the working and protection ring tunnel is established base on exit nodes.

As show in figure 7, the data traffic which enters the ring at Node A and exits at Node D follows the path A->B->C->D in normal state. When a link failure between Node B and Node C occurs, NodeB switched work ring tunnel RcW_D to opposite protection ring tunnel RaP_D same as traditionally wrapping. The different occurs in protection ring tunnel at exit node. In short wrapping protection, Rap_D will remove in Node D and deal with inner LSP label. So LSP1 will follows the path A->B->A->F->E->D when link failure between Node B and Node C when using short wrapping.

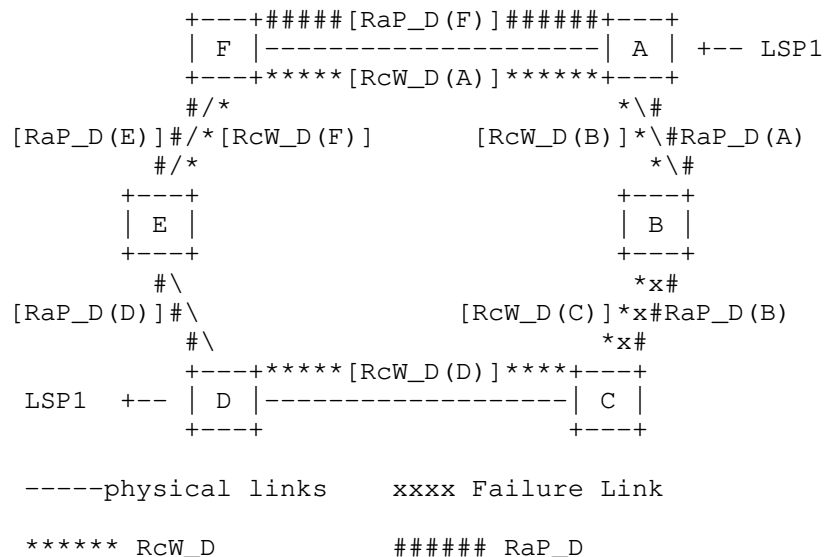


Figure 7 P2P short wrapping for link failure

2.4. P2P steering

Each working ring tunnel is associated with a protection ring tunnel in the opposite direction. Every node needs to know the ring topology by configuration or topology discovery. When the failure occurs in the ring, the nodes which detect the failure will spread the failure information in the opposite direction node by node in the ring respectively. When the node receives the message that informs the failure, it will quickly figure out the location of the fault by the topology information that is maintained by itself, so that it will determine whether the LSPs enter the ring from itself needs switch-over. If yes, it will switch the LSPs from the working ring tunnel to its protection ring tunnel.

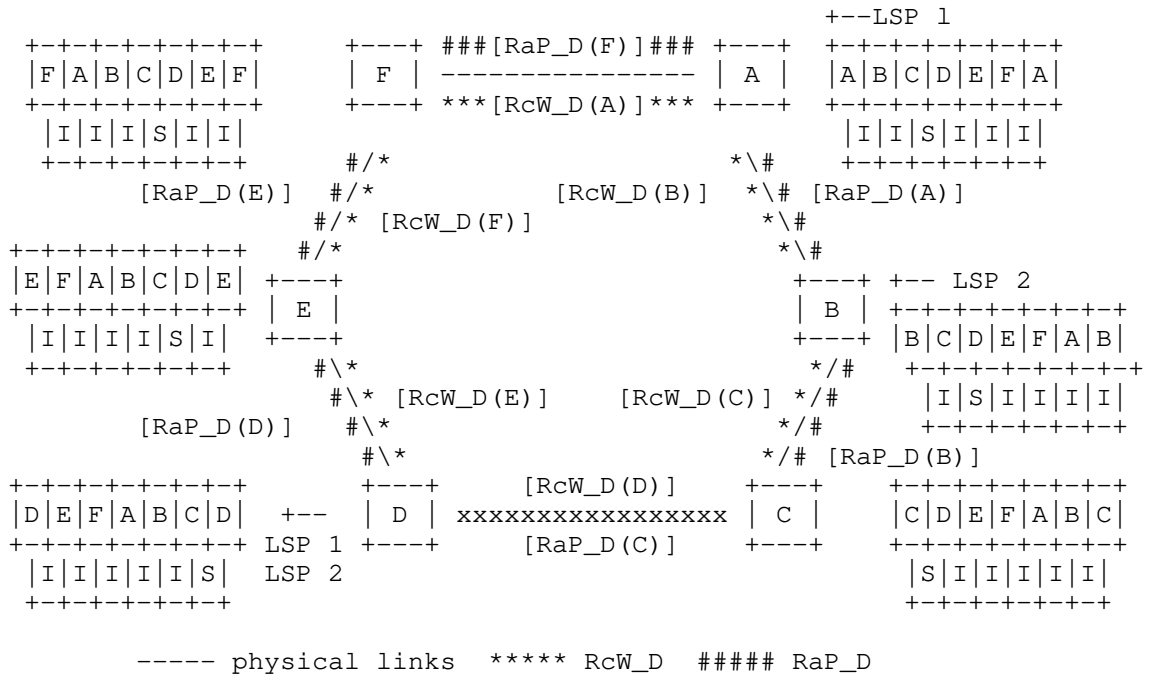


Figure 8 P2P steering operation and protection switching (1)

Steering Example is shown in Figure 8. LSP1 enters the ring from Node A while LSP2 enters the ring from Node B, and both of them have the same destination node D. As Figure 8 shows, in the normal state, LSP1 follows the path A->B->C->D, the label operation is

```
[LSP1](original data traffic carried by LSP 1
)->[RcW_D(B) | LSP1] (NodeA)->[RcW_D(C) |
LSP1] (NodeB)->[RcW_D(D) | LSP1] (NodeC)->[LSP1] ( data traffic carried
by LSP 1) . LSP2 goes through the path B->C->D, the label operation
is [LSP2]->[RcW_D(C) | LSP2] (NodeB)->[RcW_D(D) | LSP2] (NodeC)-> [LSP2] (
data traffic carried by LSP 1) .
```

If the link between C and D breaks down, as Figure 8 shows, according to the fault detection function of each link, Node D will find out that there is a failure in the link between C and D, and it will update the link state of its ring topology, changing the link state between C and D from normal to fault, as Figure 8 shows. In the direction that goes away from the failure point, Node D will send the state report message to Node E, informing Node E of the fault between C and D, and E will update the link state of its ring topology, changing the link state between C and D from normal to fault. In this manner, the state report message is sent node by node in the clockwise direction. Similar to Node D, Node C will spread the failure information in the anti-clockwise direction.

Until Node A updates the link state of its ring topology and be aware of there is a fault within its working path, it can reach the conclusion that the anticlockwise path from A to D is working all right, and thus Node A will switch the LSP1 operation to the anticlockwise ring tunnel.

```
LSP1 will follow the path A->F->E->D, the label operation is
[LSP1](original data traffic carried by LSP 1 )->[RaP_D(F) |
LSP1] (NodeA)->[RaP_D(E) | LSP1] (NodeF)->[RaP_D(D) | LSP1] (NodeE)->[LSP1]
( data traffic carried by LSP 1) .
```

The same also apply to the operation of LSP2. When Node B updates the link state of its ring topology, and finds out the working path fault, it will stop sending the LSP2 operation in the clockwise direction and switch the LSP2 to the anticlockwise protection tunnel. LSP2 goes through the path B->A->F->E->D, and the label operation is [LSP2](original data traffic carried by LSP 2)-> [RaP_D(A) | LSP2] (NodeB)->[RaP_D(F) | LSP2] (NodeA)->[RaP_D(E) | LSP2] (NodeF)->[RaP_D(D) | LSP2] (NodeE)->[LSP2] (data traffic carried by LSP 2) .

Assume that the ring between A and B breaks down, as Figure 9 shows. Like above, Node B will find out that there is a fault in the link between A and B, and it will update the link state of its ring topology, changing the link state between A and B from normal to fault. The state report message is sent node by node in the clockwise direction, informing every node that there is a fault between node A and B, so that every node updates the link state of its ring topology. Node A will find out a fault in the working path

of LSP1, and switch LSP1 to the protection Ring tunnel, while Node B will find out the LSP2 working path is all right and there is no need for switching.

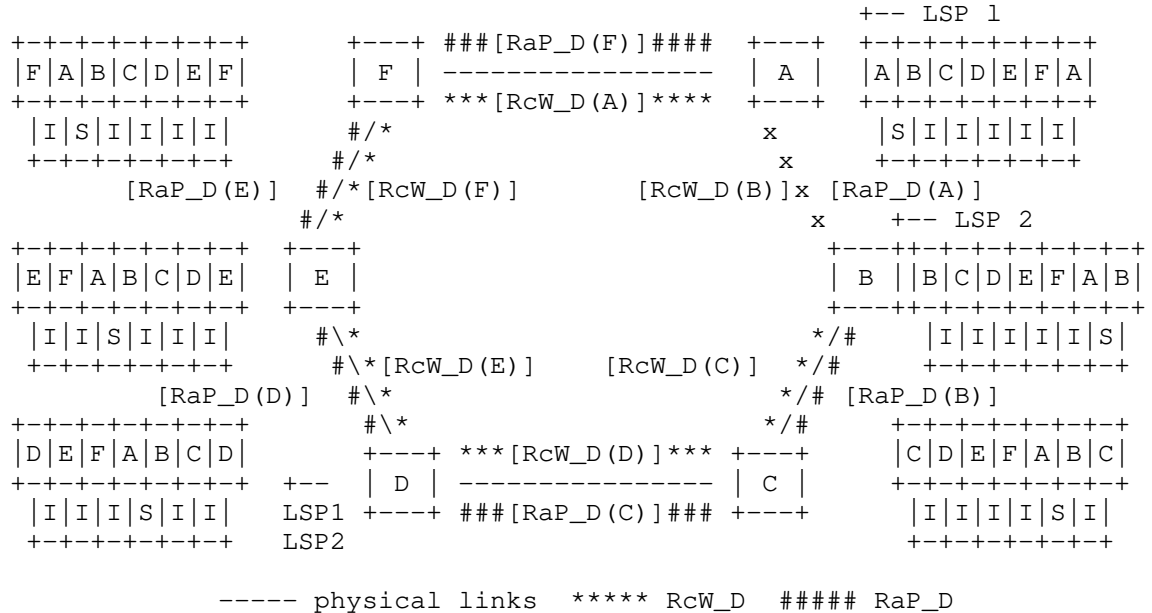


Figure 9 the P2P steering operation and protection switching (2)

2.5. P2P wrapping for Interconnected Rings

2.5.1. Interconnected ring topology

Interconnected ring topology is often used in MPLS-TP networks. There are two typical interconnected ring topologies that will be addressed in this document.

1) Single-node interconnected rings

In single-node interconnected rings, the connection between two rings is through a single node. As the interconnection node may cause a single point of failure, this topology should be avoided in real networks;

2) Dual-node interconnected rings

In dual-node interconnected rings, the connection between two rings is through two nodes. The two interconnection nodes belong to both

interconnected rings. This topology can recover from one interconnection node failure.

2.5.1.1. Single-node interconnected rings

Figure 10 shows the topology of single-node interconnected rings. Node C is interconnection node between Ring1 and Ring2.

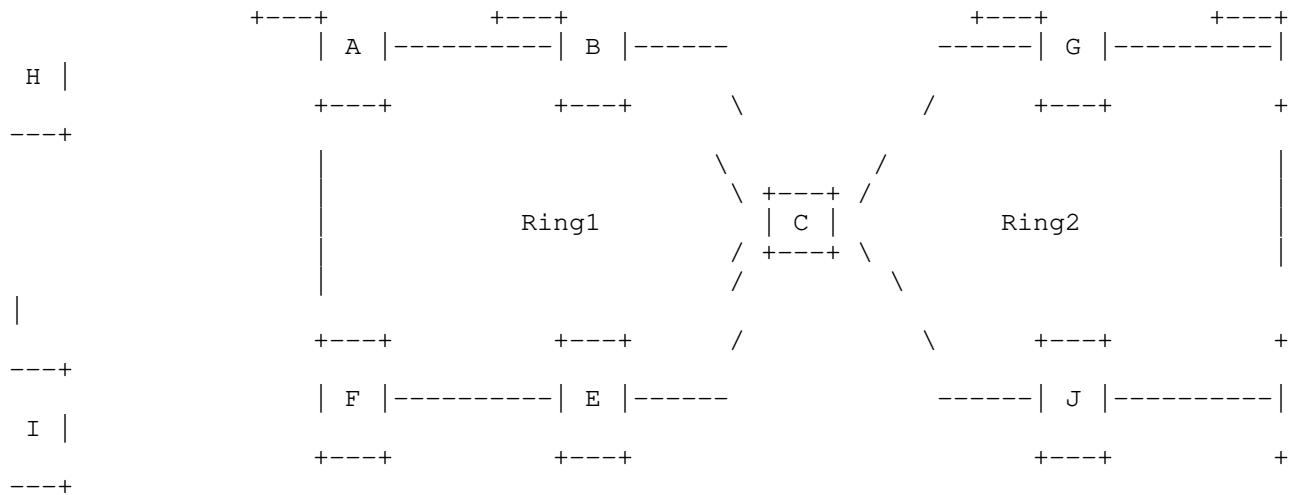


Figure 10 Single-node interconnected rings

2.5.1.2. Dual-node interconnected rings

Figure 11 shows the topology of dual-node interconnected rings. Node C and Node D are interconnection nodes between Ring1 and Ring2.

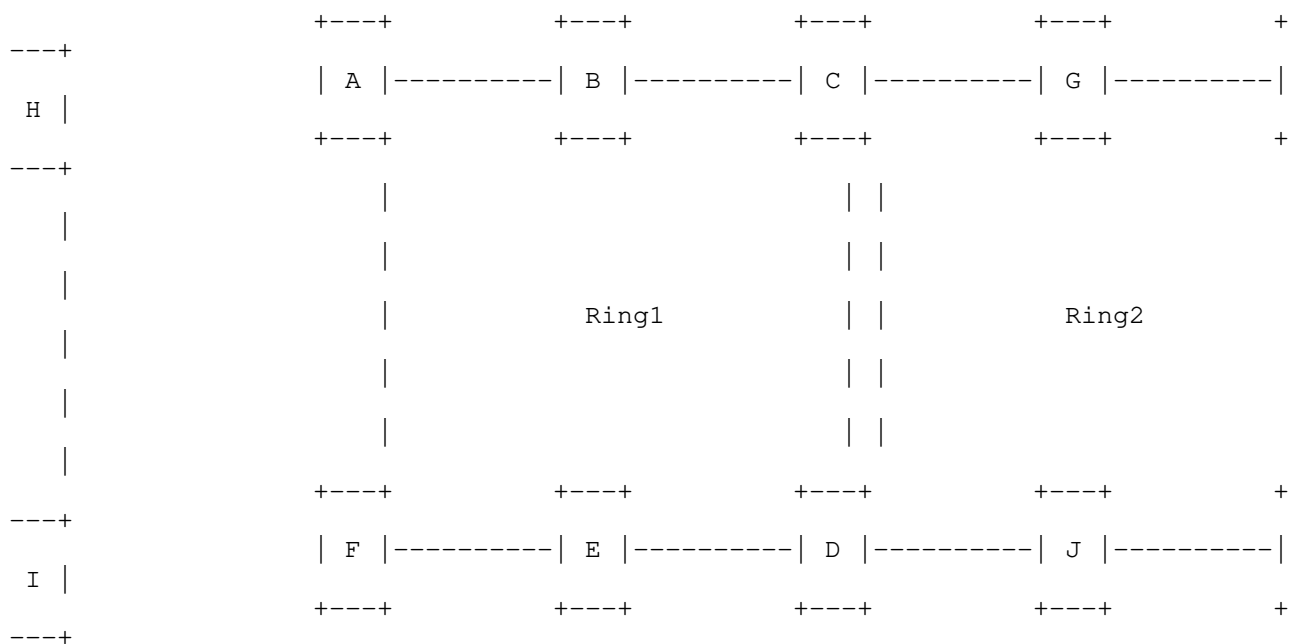


Figure 11 Dual-node interconnected rings

2.5.2. Interconnected ring protection scheme

Cheng, et al.

Expires January 09, 2014

[Page 15]

2.5.2.1. Introduction

- Interconnected rings can be regarded as two independent rings. Each ring runs protection switching independently. Failure in one ring only triggers protection switching in itself and does not affect the other ring. Protection switch in a single ring is same as which described in section 3 Shared ring protection for P2P.
- The service LSPs that traverse the interconnected rings via the interconnection nodes must use different ring tunnels in different rings. The ring tunnel used in the source ring will be removed, and the ring tunnel of destination ring will be added in interconnection nodes.
- For protected interconnection node in dual-node interconnected ring, the service LSPs in the interconnection nodes should use the same MPLS label. So any interconnection node can terminate source ring tunnel and push destination ring tunnel according to service LSP label.
- Two interconnection nodes can be managed as a virtual interconnection node group. Each ring should assign ring tunnels to the virtual interconnection node group. The interconnection nodes in the group should terminate the working ring tunnel in each ring. Protection ring tunnel is a open ring to switch with the working ring tunnel at the nodes which detect the fault and end at the egress node.
- When the service traffic passes through the interconnection node, the direction of the working ring tunnels in each ring for this service traffic should be the same. For example, if the working ring tunnel follows the clockwise direction in Ring1, the working ring tunnel for the same service traffic in Ring2 also follows the clockwise direction when the service leaves Ring1 and enters Ring2.

2.5.2.2. Ring tunnels of interconnected rings

The same ring tunnels as described in 2.1.1 are used in each ring of the interconnected rings. Besides, ring tunnels to the virtual interconnection node group will be established by each ring of the interconnected rings, i.e.:

- one clockwise working ring tunnel to the virtual interconnection node group;
- one anticlockwise protection ring tunnel to the virtual interconnection node group,

- one anticlockwise working ring tunnel to the virtual interconnection node group;
- one clockwise protection ring tunnel to the virtual interconnection node group.

These ring tunnel will terminated at all nodes in virtual interconnection node group.

All the ring tunnels established in Ring1 in Figure 11 is provided as follows:

To Node A: R1cW_A, R1aW_A, R1cP_A, R1aP_A;

To Node B: R1cW_B, R1aW_B, R1cP_B, R1aP_B;

To Node C: R1cW_C, R1aW_C, R1cP_C, R1aP_C;

To Node D: R1cW_D, R1aW_D, R1cP_D, R1aP_D;

To Node E: R1cW_E, R1aW_E, R1cP_E, R1aP_E;

To Node F: R1cW_F, R1aW_F, R1cP_F, R1aP_F;

To the virtual interconnection node group (including Node F and Node A): R1cW_F&A, R1aW_F&A, R1cP_F&A, R1aP_F&A;

All the ring tunnels established in Ring2 in Figure 11 is provided as follows:

To Node A: R2cW_A, R2aW_A, R2cP_A, R2aP_A;

To Node F: R2cW_F, R2aW_F, R2cP_F, R2aP_F;

To Node G: R2cW_G, R2aW_G, R2cP_G, R2aP_G;

To Node H: R2cW_H, R2aW_H, R2cP_H, R2aP_H;

To Node I: R2cW_I, R2aW_I, R2cP_I, R2aP_I;

To Node J: R2cW_J, R2aW_J, R2cP_J, R2aP_J;

To the virtual interconnection node group (including Node F and Node A): R2cW_FandA, R2aW_FandA, R2cP_FandA, R2aP_FandA;

2.5.2.3. Interconnected ring switch mechanism

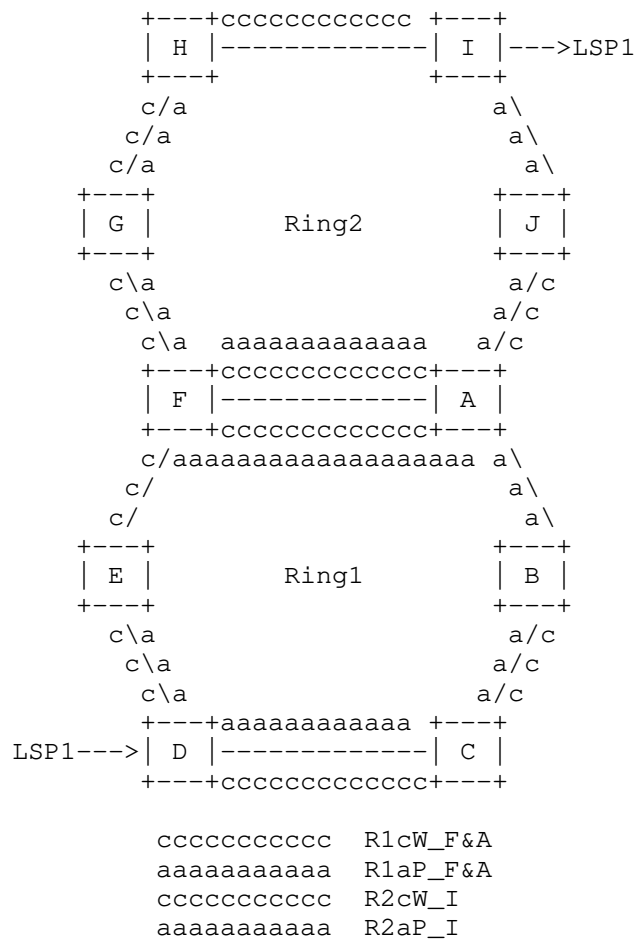


Figure 12 Ring tunnels for the interconnected rings

As shown in Figure 12, for the service traffic LSP1 which enters Ring1 at Node D and leaves Ring1 at Node F and continues to enter Ring2 at Node F and leaves Ring2 at Node I, the protection scheme is described below.

In normal state, LSP1 follows R1cW_F&A in Ring1 and R2cW_I in Ring2. The label used for the working ring tunnel R1cW_F&A in Ring1 is popped and the label used for the working ring tunnel R2cW_I will be pushed based the inner label lookup at the interconnection node F. The working path that the service traffic LSP1 follows is:
LSP1->R1cW_F&A (D->E->F)->R2cW_I (F->G->H->I)->LSP1.

In case of link failure, for example, when a failure occurs on the link between Node F and Node E, Node F and E will detect the failure and execute protection switching as described in 2.2.1.1. The path that the service traffic LSP1 follows after switching change to
 LSP1->R1cW_F&A(D->E)->R1aP_F&A(E->D->C->B->A->F)->R1cW_F(F)
 ->R2cW_I(F->G->H->I)->LSP1.

In case of non interconnection node failure, for example, when the failure occurs at Node E in Ring1, Node F and E will detect failure and execute protection switching as described in 2.2.1.2. The path that the service traffic LSP1 follows after switching becomes:
 LSP1->R1cW_F&A(D)->R1aP_F&A(D->C->B->A->F)->
 R1cW_F(F)->R2cW_I(F->G->H->I).

In case of interconnection node failure, for example, when failure occurs at the interconnection Node F. Node E and A in Ring1 will detect the failure, and execute protection switching as described in 2.2.1.2. Node G and A in Ring2 will also detects the failure, and execute protection switching. The path that the service traffic LSP1 follows after switching is:
 LSP1->R1cW_F&A(D->E)->R1aP_F&A(E->D->C->B->A)->R1cW_A(A)
 ->R2aP_I(A->J->I)->LSP1.

2.5.2.4. Interconnected ring topology detection mechanism

As show in Figure 13, the service traffic LSP1 traverses A->B-C in Ring1 and C->G->H->I in Ring2. Node C and Node D is the interconnection node. When both the link between Node C and Node G and the link between Node C and Node D fail, ring tunnel from Node C to Node I in Ring 2 becomes unreachable. However, Node D is still available, by which LSP1 can still reach Node I.

In order to do so, the interconnection nodes need to know the ring topology in each ring independently so that they can judge whether a node is reachable. The judgment is based on the knowledge of ring topology and the fault location as described in section 3.4. The ring topology can be obtained by NMS or topology discovery mechanisms. The fault location can be obtained by spreading the fault information around the ring. The nodes which detect the failure will spread the fault information in the opposite direction node by node in the ring respectively. When the interconnection node receives the message that informs the failure, it will quickly figure out the location of the fault by the topology information that is maintained by itself and determine whether the LSPs enter the ring from itself can reach the destination. If the destination node is reachable, the LSP will exit the source ring and enter the destination ring. If the destination node is not reachable, the LSP will switch to the anticlockwise protection ring tunnel.

In Figure 13 Node C judges the ring tunnel to Node I is unreachable, the service traffic LSP1 of which the destination node on the ring tunnel is Node I should switch to the protection LSP (R1aP_C&D) so that the service traffic LSP1 traverses the interconnected rings at Node D. Node D will remove the ring tunnel label of Ring1 and add ring tunnel label of Ring2.

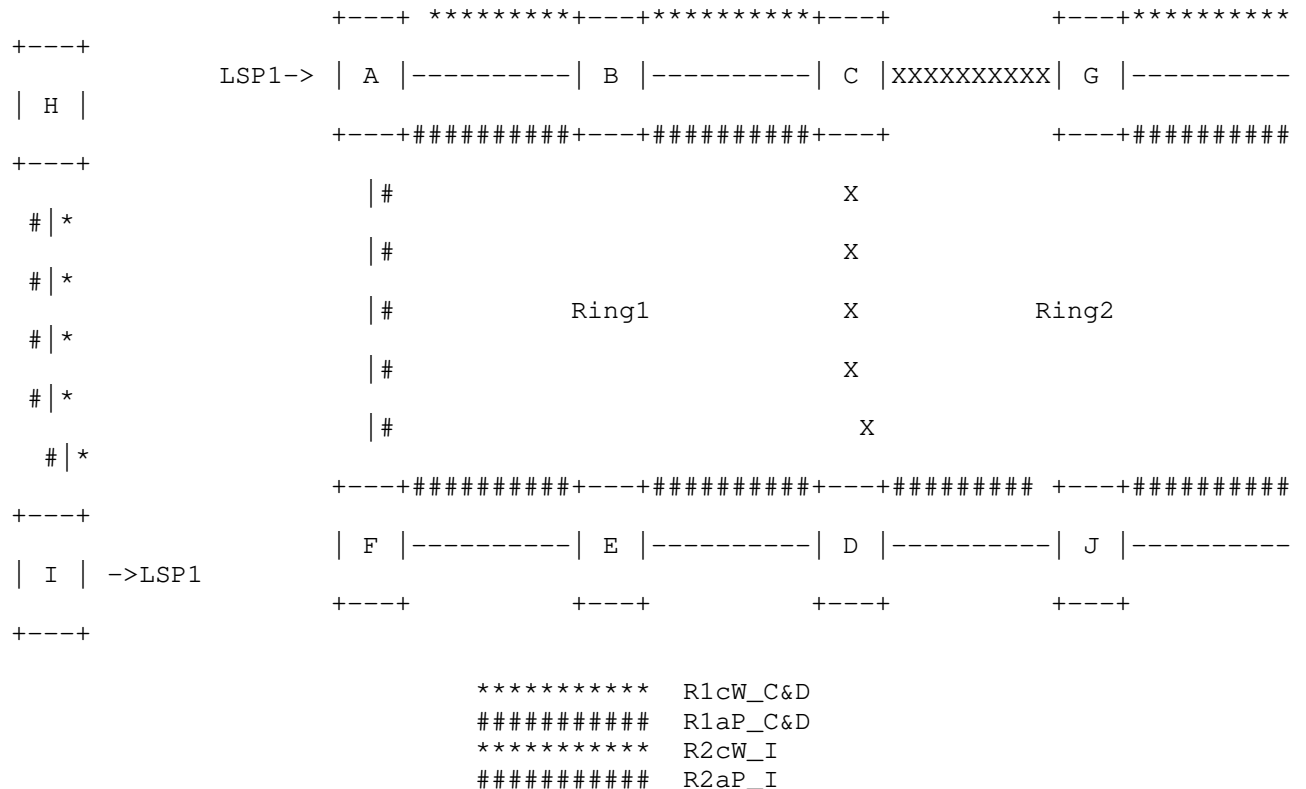


Figure 13 interconnected ring

3. Coordination protocol

TBD

4. Conclusions

TBD

5. IANA Considerations

None

6. Security Considerations

TBD

7. References

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

- [RFC3031] Rosen, E., Viswanathan, A., and R. Callon, "Multiprotocol Label Switching Architecture", RFC 3031, January 2001.
- [RFC5654] Niven-Jenkins, B., Brungard, D., Betts, M., Sprecher, N., and S. Ueno, "Requirements of an MPLS Transport Profile", RFC 5654, September 2009.
- [RFC6371] Busi, I. and D. Allan, "Operations, Administration, and Maintenance Framework for MPLS-Based Transport Networks", RFC 6371, September 2011.

Authors' Addresses

Weiqiang Cheng
China Mobile
No.32 Xuanwumen West Street
Beijing 100053
China

Email: chengweiqiang@chinamobile.com

Lei Wang
China Mobile
No.32 Xuanwumen West Street
Beijing 100053
China

Email: Wangleiyj@chinamobile.com

Han Li
China Mobile
No.32 Xuanwumen West Street
Beijing 100053
China

Email: Lihan@chinamobile.com

Kai Liu
Huawei Technologies Co., Ltd.
Huawei base, Bantian, Longgang District
Shenzhen 518129
China

Email: alex.liukai@huawei.com

Jia He
Huawei Technologies Co., Ltd.
Huawei base, Bantian, Longgang District
Shenzhen 518129
China

Email: hejia@huawei.com

Fang Li
China Academy of Telecommunication Research, MIIT., China
No.52 Huayuan Street
Beijing 100191
China

Email: lifang@rit.cn

Jian Yang
ZTE Corporation P.R.China
ZTE Industrial Zone, Liuxian Road
Shenzhen 518055
China

Email: yang.jian90@zte.com.cn

Junfang Wang
Fiberhome Telecommunication Technologies Co., LTD.
No.5, Dongxin Lu
Wuhan 430073
China

Email: wjf@fiberhome.com.cn

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: January 16, 2014

Z. Cui
R. Winter
NEC
L. Zheng
M. Chen
Huawei Technologies Ltd.
July 15, 2013

ICC Based TLVs for MPLS-TP OAM Functions
draft-cui-mps-tp-oam-tlv-icc-00

Abstract

This document specifies identifier TLVs for the Multiprotocol Label Switching Transport Profile (MPLS-TP). Several IP-compatible TLVs have already been defined. This document extends the existing set of identifier TLVs based on MPLS-TP identifiers following ITU-T conventions.

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 January 16, 2014.

Copyright Notice

Copyright (c) 2013 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	2
1.1. Terminology	2
1.2. Requirements notation	3
2. ICC_Operator_ID-based TLVs and sub-TLVs Format	3
2.1. ICC_Operator_ID-based Source ID TLV	3
2.2. ICC_Operator_ID-based Destination ID TLV	4
2.3. ICC_Operator_ID-based Source MEP ID TLV	4
2.4. ICC_Operator_ID TLV	5
2.5. ICC_Operator_ID-based Static LSP sub-TLV	5
2.6. ICC_Operator_ID-based Static Pseudowire sub-TLV	6
3. Extension TLVs for MPLS-TP OAM Functions	7
3.1. Proactive Connectivity Verification, Continuity Check, and Remote Defect Indication	7
3.2. On-demand Connectivity Verification and Route Tracing . .	8
3.3. MPLS Fault Management	8
3.4. Lock Instruct and Loopback	8
3.5. Packet Loss and Delay Measurement	9
4. Security Considerations	9
5. IANA Considerations	9
5.1. New ICC-based Source/Destination ID TLVs	9
5.2. New ICC-based Source MEP-ID TLV	9
5.3. New ICC_Operator_ID TLV	10
6. New ICC-based Static LSP and PW TLVs	10
7. Normative References	10
Authors' Addresses	11

1. Introduction

Several MPLS-TP OAM functions such as the ones defined in [RFC6426], [RFC6427], [RFC6428], [RFC6435] utilize OAM PDUs which carry Identifiers (ID) based on IP/MPLS conventions including IP addresses and AS numbers.

However, in ITU-T specified transport networks, the ITU Carrier Code (ICC) is traditionally used to identify a carrier/service provider. This document defines the corresponding TLVs based on IDs following ITU-T conventions [RFC6923].

1.1. Terminology

OAM: Operations, Administration, and Maintenance

TLV: Type Length Value

ITU-T: ITU Telecommunication Standardization Sector

CC: Country Code

ICC: ITU Carrier Code

MPLS: Multiprotocol Label Switching

LSP: Label Switched Path

PW: Pseudowire

MEG: Maintenance Entity Group

UMC: Unique MEG ID Code

MEP: Maintenance Entity Group End Point

MIP: Maintenance Entity Group Intermediate Point

1.2. Requirements notation

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].

2. ICC_Operator_ID-based TLVs and sub-TLVs Format

This document defines ICC_Operator_ID-based TLVs and sub-TLVs to be used in place of exiting TLVs based on IP conventions.

2.1. ICC_Operator_ID-based Source ID TLV

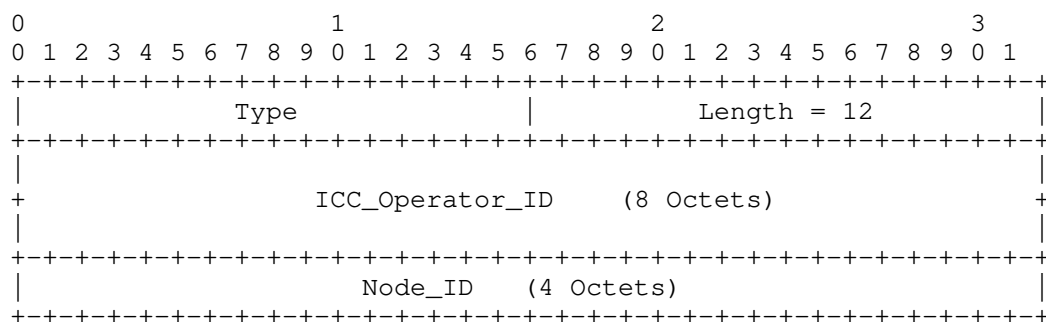


Figure 1: ICC_Operator_ID-based Source ID TLV Format

The encoding of the ICC_Operator_ID-based Source ID TLV is specified in the figure above.

Type: TBD1 (two octets).

Length: The length field is two octets and indicates the length of the TLV value field. This TLV has a fixed length and the value is always 12.

ICC Operator ID: The ICC Operator ID field is 8 octets and is encoded as specified in [RFC6923].

Node ID: The node ID field is a 4 octet field and is encoded as specified in [RFC6370].

This TLV does not carry sub-TLVs.

2.2. ICC_Operator_ID-based Destination ID TLV

The ICC_Operator_ID-based Destination Identifier TLV has the same format as ICC_Operator_ID-based Source Identifier TLV that is specified in the figure above (see Figure 1).

The TLV Type is TBD2.

2.3. ICC_Operator_ID-based Source MEP ID TLV

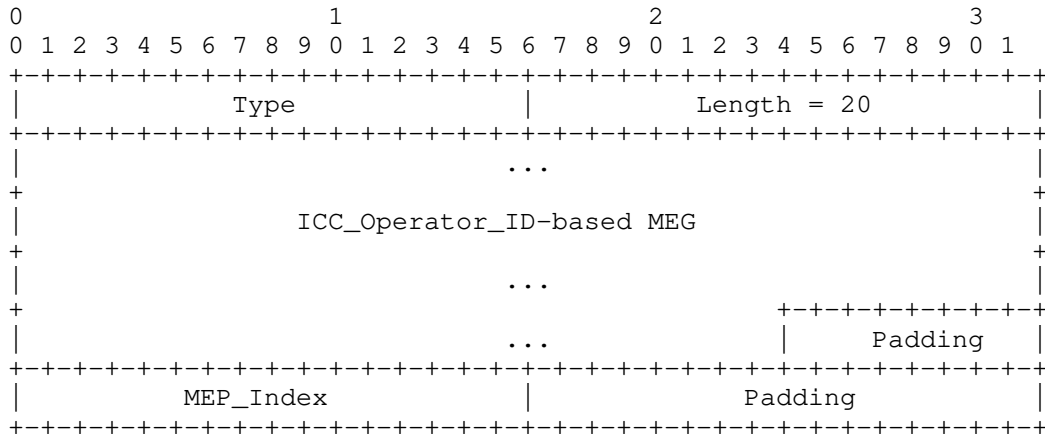


Figure 2: ICC_Operator_ID-based Source MEP ID TLV Format

The encoding of the ICC_Operator_ID-based Source MEP ID TLV is specified in the figure above.

Type: TBD3 (two octets).

Length: The length field is two octets in length and indicates the length of the TLV value part in octets including the padding. This TLV has a fixed length and the value is always 20.

ICC_Operator_ID-based MEG: The ICC_Operator_ID-based MEG field is 15 octets and is encoded as specified in [RFC6923].

Value: The value field is encoded by appending a 16-bit MEP index as defined in [RFC6923].

Padding: the padding field is used to align the field with a 4-octet boundary. The length of the padding MUST be included in the length field.

This TLV does not carry sub-TLVs.

2.4. ICC_Operator_ID TLV

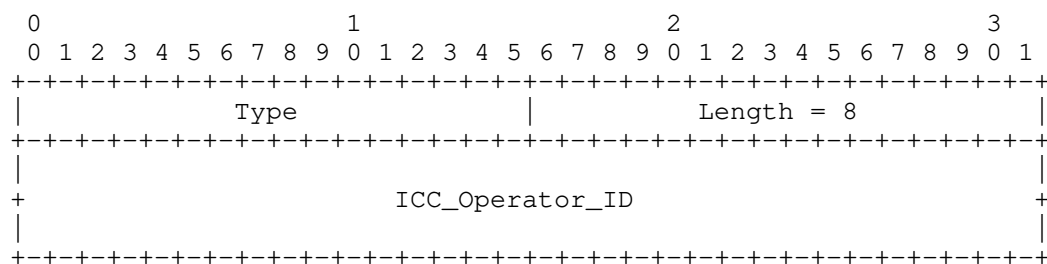


Figure 3: ICC_Operator_ID TLV Format

The encoding of the ICC_Operator_ID TLV is specified in the figure above.

Type: TBD4 (two octets).

Length: The length field is two octets in length and indicates the length of the TLV value part in octets. This TLV has a fixed length and the value is always 8.

ICC_Operator_ID: The ICC_Operator_ID field is 8 octets and is encoded as specified in [RFC6923].

This TLV does not carry sub-TLVs.

2.5. ICC_Operator_ID-based Static LSP sub-TLV

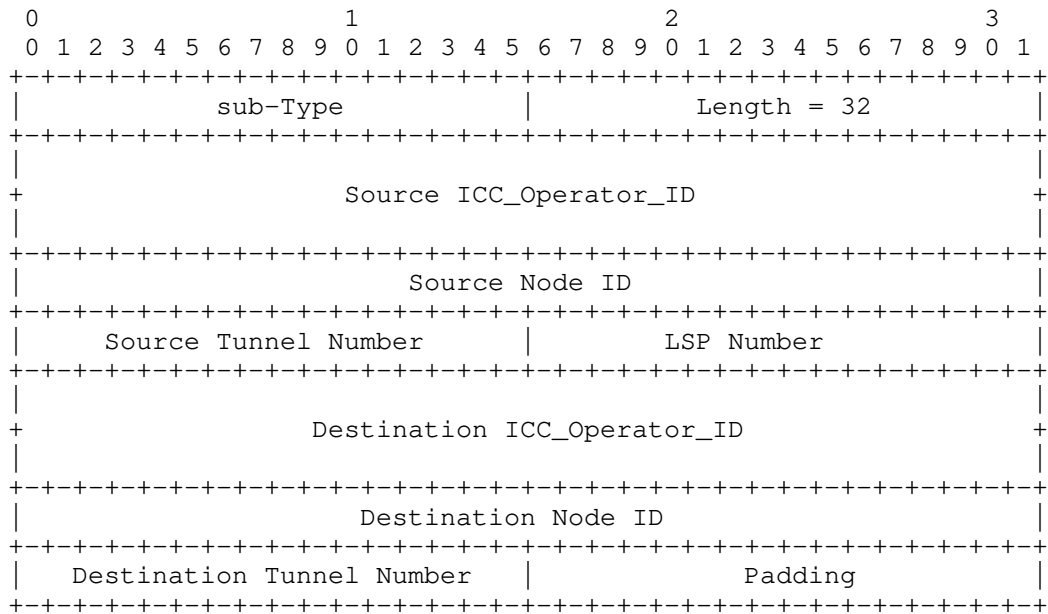


Figure 4: ICC_Operator_ID-based Static LSP sub-TLV Format

The encoding of the ICC_Operator_ID-based Static LSP sub-TLV is specified in the figure above.

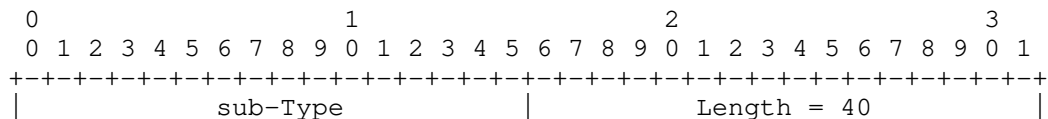
sub-Type: TBD5 (two octets).

Length: The length field is two octets in length and indicates the length of the TLV value part in octets including the padding. This TLV has a fixed length and the value is always 32.

The ICC_Operator_ID fields are defined in [RFC6923]. The Node_ID, Tunnel Number and LSP Number are defined in [RFC6370]. The Source ICC_Operator_ID and Destination ICC_Operator_ID MAY be set to zero in case global uniqueness is not required.

Padding: the padding field is used to align the field with a 4-octet boundary. The length of the padding MUST be included in the length field.

2.6. ICC_Operator_ID-based Static Pseudowire sub-TLV



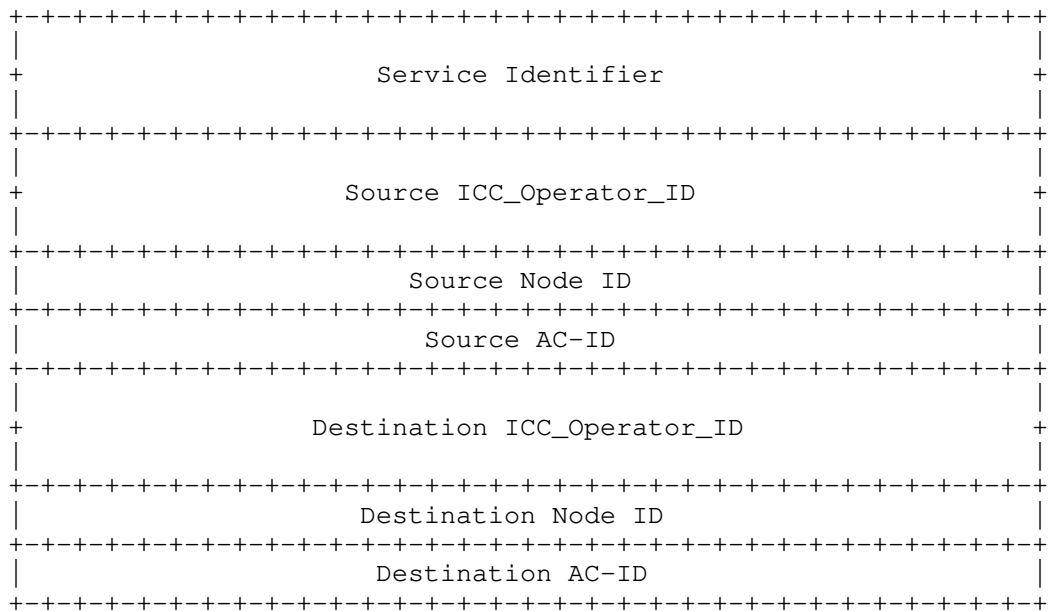


Figure 5: ICC_Operator_ID-based Static Pseudowire sub-TLV Format

The encoding of the ICC_Operator_ID-based Static Pseudowire sub-TLV is specified in the figure above.

sub-Type: TBD6 (two octets).

Length: The length field is two octets in length and indicates the length of the TLV value part in octets. This TLV has a fixed length and the value is always 40.

The Service Identifier field is defined in [RFC6426]. The ICC_Operator_ID fields are defined in [RFC6923]. The Node_ID fields are defined in [RFC6370]. The AC-ID fields are defined in [RFC5003].

3. Extension TLVs for MPLS-TP OAM Functions

This document specifies the extension for MPLS-TP OAM functions using ICC-based TLVs, but does not change the behavior of existing OAM functions.

3.1. Proactive Connectivity Verification, Continuity Check, and Remote Defect Indication

[RFC6428] defines the format of an MPLS-TP CV Message to allow the MEPS to proactively monitor the liveliness and connectivity of a

transport path (LSP, PW, or a Section) between them. [RFC6428] defines three types of Source MEP-ID TLV for MPLS-TP Connectivity Verification such as section MEP-ID, LSP MEP-ID and PW End Point MEP-ID.

When sending MPLS-TP CV Message on ICC based network, the Source MEP-ID TLV based on ITU-T conventions (see Section 2.3) MUST be used in the MPLS-TP CV Message, with no distinction between sections, LSPs and Pseudowires.

3.2. On-demand Connectivity Verification and Route Tracing

[RFC6426] specifies an on-demand monitoring mechanism for the MPLS Transport Profile (MPLS-TP). It defines a set of Global_ID-based Source/Destination Identifier TLVs to identify the source or destination node of an OAM message, and also defined a set of Global_ID-based Static LSP sub-TLV and Static Pseudowire sub-TLV to identify the LSP and Pseudowire path.

When sending Non-IP-Based On-Demand CV Packet on ICC based network, the Source Identifier TLV (see Section 2.1) and Destination Identifier TLV (see Section 2.2) based on ITU-T conventions MUST be used in the On-Demand CV Packet instead of Global_ID-based Source Identifier TLV and Global_ID-based Destination Identifier TLV. The Static LSP sub-TLVs and Static Pseudowire sub-TLVs also Must be replaced by ICC_Operator_ID-based Static LSP sub-TLVs (see Section 2.5) and ICC_Operator_ID-based Static Pseudowire sub-TLVs (see Section 2.6).

3.3. MPLS Fault Management

[RFC6427] specifies a Fault Management messages to indicate service disruptive conditions for MPLS-based transport network Label Switched Paths, the Fault Management message include a Global_ID TLV.

When sending Non-IP-Based Fault Management messages on ICC based network, the ICC_Operator_ID (see Section 2.4) MUST be used in the Fault Management messages instead of the Global_ID TLV based on IP conventions.

3.4. Lock Instruct and Loopback

[RFC6435] specifies the Lock Instruct message include Source MEP-ID TLV to allow the MEPs to lock transport path (LSP, PW, or a Section).

When sending Non-IP-Based Fault Management messages on ICC based network, the Source MEP-ID TLV based on ITU-T conventions (see Section 2.3) MUST be used in the Lock Instruct message instead of TLVs based on IP conventions.

3.5. Packet Loss and Delay Measurement

[RFC6374] defined the Loss Measurement Message used to measure packet loss ratio and packet delay. This document does not define new TLV for it, because no Global_ID TLV is used in this function.

4. Security Considerations

TBD

5. IANA Considerations

5.1. New ICC-based Source/Destination ID TLVs

Section 2 defines two new TLV types for use with on-demand CV ([RFC6426]).

IANA is requested to assign the following TLV types from the "Multiprotocol Label Switching Architecture (MPLS) Label Switched Paths (LSPs) Ping Parameters - TLVs" registry, "TLVs and sub-TLVs" sub-registry.

Type	Length	Value Field
-----	-----	-----
TBD1	12	ICC_Operator_ID-based Source ID TLV
TBD2	12	ICC_Operator_ID-based Destination ID TLV

5.2. New ICC-based Source MEP-ID TLV

Section 2 defines a new TLV types for use with proactive cc-cv-rdi ([RFC6426]).

IANA is requested to assign the following TLV types from the "Multiprotocol Label Switching Architecture (MPLS) Label Switched Paths (LSPs) Ping Parameters - TLVs" registry, "TLVs and sub-TLVs" sub-registry.

Type	Length	Value Field
-----	-----	-----
TBD3	20	ICC_Operator_ID-based Source MEP-ID TLV

5.3. New ICC_Operator_ID TLV

Section 2 defines a new TLV type for use with MPLS-TP Fault Management ([RFC6427]).

IANA is requested to assign the following TLV types from the "Multiprotocol Label Switching Architecture (MPLS) Label Switched Paths (LSPs) Ping Parameters - TLVs" registry, "TLVs and sub-TLVs" sub-registry.

Type	Length	Value Field
-----	-----	-----
TBD4	8	ICC_Operator_ID TLV

6. New ICC-based Static LSP and PW TLVs

Section 2 defines two new sub-TLV types for use with on-demand CV([RFC6426]).

IANA is requested to assign the following sub-TLV types from the "Multiprotocol Label Switching Architecture (MPLS) Label Switched Paths (LSPs) Ping Parameters - TLVs" registry, "TLVs and sub-TLVs" sub-registry.

sub-Type	Length	Value Field
-----	-----	-----
TBD5	32	ICC_Operator_ID-based Static LSP
TBD6	40	ICC_Operator_ID-based Static PW

7. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC5003] Metz, C., Martini, L., Balus, F., and J. Sugimoto, "Attachment Individual Identifier (AII) Types for Aggregation", RFC 5003, September 2007.
- [RFC6370] Bocci, M., Swallow, G., and E. Gray, "MPLS Transport Profile (MPLS-TP) Identifiers", RFC 6370, September 2011.
- [RFC6374] Frost, D. and S. Bryant, "Packet Loss and Delay Measurement for MPLS Networks", RFC 6374, September 2011.

- [RFC6426] Gray, E., Bahadur, N., Boutros, S., and R. Aggarwal, "MPLS On-Demand Connectivity Verification and Route Tracing", RFC 6426, November 2011.
- [RFC6427] Swallow, G., Fulignoli, A., Vigoureux, M., Boutros, S., and D. Ward, "MPLS Fault Management Operations, Administration, and Maintenance (OAM)", RFC 6427, November 2011.
- [RFC6428] Allan, D., Swallow Ed. , G., and J. Drake Ed. , "Proactive Connectivity Verification, Continuity Check, and Remote Defect Indication for the MPLS Transport Profile", RFC 6428, November 2011.
- [RFC6435] Boutros, S., Sivabalan, S., Aggarwal, R., Vigoureux, M., and X. Dai, "MPLS Transport Profile Lock Instruct and Loopback Functions", RFC 6435, November 2011.
- [RFC6923] Winter, R., Gray, E., van Helvoort, H., and M. Betts, "MPLS Transport Profile (MPLS-TP) Identifiers Following ITU-T Conventions", RFC 6923, May 2013.

Authors' Addresses

Zhenlong Cui
NEC

Email: c-sai@bx.jp.nec.com

Rolf Winter
NEC

Email: Rolf.Winter@neclab.eu

Lianshu Zheng
Huawei Technologies Ltd.

Email: vero.zheng@huawei.com

Mach(Guoyi) Chen
Huawei Technologies Ltd.

Email: mach.chen@huawei.com

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: September 29, 2013

A. D'Alessandro
Telecom Italia
J. Ryoo
ETRI
H. van Helvoort
Huawei Technologies
March 28, 2013

Supporting the Exercise command for PSC linear protection protocol
draft-dj-mpls-tp-exer-psc-01

Abstract

This draft indicates how IETF RFC6378 could be modified to address the Exercise function.

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 September 29, 2013.

Copyright Notice

Copyright (c) 2013 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	2
2. Updates to the PSC RFC	3
2.1. Updates to Section 2.1. Acronyms	3
2.2. Updates to Section 3.1. Local Request Logic	3
2.3. Update to Section 3.2. Remote Requests	3
2.4. Updates to Section 3.6. PSC Control States	3
2.5. Updates to Section 4.2.2. PSC Request Field	4
2.6. Updates to Section 4.3.2. Priority of Inputs	4
2.7. Updates to Section 4.3.3. Operation of PSC States	4
2.7.1. Updates to Section 4.3.3.1. Normal State	4
2.7.2. Updates to Section 4.3.3.6. Do-not-Revert State	4
2.7.3. New subsection for Exercise State	4
2.8. Updates to Appendix A. PSC State Machine Tables	6
2.9. Updates to Appendix B. Exercising the Protection Domain	9
3. IANA Considerations	9
4. Security Considerations	9
5. Acknowledgements	9
6. References	9
6.1. Normative References	9
6.2. Informative References	10
Authors' Addresses	10

1. Introduction

Exercise is a command to test if the PSC communication is operating correctly. More specifically, the Exercise is to test and validate the linear protection mechanism and PSC protocol including the aliveness of the Local Request logic, the PSC state machine and the PSC message generation and reception, and the integrity of the protection path, without triggering the actual traffic switching. It is used while the working path is either carrying the traffic or not. It is lower priority than any "real" switch request. It is only valid in bidirectional switching, since this is the only place where one can get a meaningful test by looking for a response.

This command is documented in R84 of [RFC5654] and it has been identified as a requirement in the ITU's liaison statement "Liaison Statement: Recommendation ITU-T G.8131/Y.1382 revision - Linear protection switching for MPLS-TP networks " [LIAISON1205] and "Recommendation ITU-T G.8131 revision - Linear protection switching for MPLS-TP networks [LIAISON1234]. This draft is created as an attempt to align PSC behaviour and functionalities to meet IETF and ITU-T MPLS Transport Profile requirements.

2. Updates to the PSC RFC

This section describes the changes required to cover the exercise functionality to the PSC protocol defined in [RFC6378]

2.1. Updates to Section 2.1. Acronyms

The following text should be added in Section 2.1 in [RFC6378]:

EXER Exercise
RR Reverse Request

2.2. Updates to Section 3.1. Local Request Logic

EXER should be included as an operator command.

The following text should be added:

- o Exercise (EXER) - Exercise is a command to test if the PSC communication is operating correctly. It is lower priority than any "real" switch request. It is only valid in bidirectional switching, since this is the only place where one can get a meaningful test by looking for a response.

2.3. Update to Section 3.2. Remote Requests

The following text should be added:

- o Remote EXER - indicates that the remote end point is operating under an operator command to validate the protection mechanism and PSC protocol including the aliveness of the Local Request logic, the PSC state machine and the PSC message generation and reception, and the integrity of the protection path, without triggering the actual traffic switching. The valid response to EXER message will be an RR with the corresponding FPath and Path numbers. The near end will signal a Reverse Request (RR) only in response to an EXER command from the far end.

When Exercise commands are input at both ends, an EXER, instead of RR, is transmitted from both ends.

2.4. Updates to Section 3.6. PSC Control States

The following text should be added:

- o Exercise state - The operator has issued the Exercise command to test and validate the protection mechanism and PSC protocol including the integrity of the protection path, without triggering the actual traffic switching.

2.5. Updates to Section 4.2.2. PSC Request Field

The following PSC Requests should be added to PSC Request field:

(3) Exercise - indicates that the transmitting end point is exercising the protection channel and mechanism. FPath and Path are set to the same value of the NR, RR or DNR request that EXER replaces.

(2) Reverse Request - indicates that the transmitting end point is responding to an EXER command from the far end. FPath and Path are set to the same value of the NR, RR or DNR request that EXER replaces.

2.6. Updates to Section 4.3.2. Priority of Inputs

The priority of the Exercise should be inserted between the priorities of WTR Expires and No Request.

2.7. Updates to Section 4.3.3. Operation of PSC States

2.7.1. Updates to Section 4.3.3.1. Normal State

Add the following text for Section 4.3.3.1. Normal State:

- o A local Exercise input SHALL cause the LER to go into local Exercise state and begin transmission of an EXER(0,0) message.
- o A remote EXER message SHALL cause the LER to go into remote Exercise state, and transmit an RR(0,0)message.

2.7.2. Updates to Section 4.3.3.6. Do-not-Revert State

Add the following text for Section 4.3.3.6. Do-not-Revert State:

- o A local Exercise input SHALL cause the LER to go into local Exercise state and begin transmission of an EXER(0,1) message.
- o A remote EXER message SHALL cause the LER to go into remote Exercise state, and transmit an RR(0,1)message.

2.7.3. New subsection for Exercise State

Add a new sub-section, Section 4.3.3.7. Exercise State, with the following text:

In the Exercise state, the user data traffic SHALL remain on the same path as the previous state, such as Normal state or Do-Not-Revert state. The local end SHALL signal a RR message in response to a remote EXER message. When both ends are in local Exercise state, only the EXER messages are exchanged.

When in Exercise state, the following describe the reaction to local input:

- o A local Clear SHALL be ignored if in remote Exercise state. If in local Exercise state, then this input SHALL cause the LER to go into Normal state and begin transmitting NR(0,0) when the LER is configured for revertive mode. For non-revertive mode, the LER goes into DNR state and begin transmitting DNR(0,1).
- o A local Lockout of protection input SHALL cause the LER to go into local Unavailable state and begin transmission of an LO(0,0) message.
- o A local Forced Switch input SHALL cause the LER to go into local Protecting administrative state and begin transmission of an FS(1,1) message.
- o A local Signal Fail indication on the protection path SHALL cause the LER to go into local Unavailable state and begin transmission of an SF(0,0) message.
- o A local Signal Fail indication on the working path SHALL cause the LER to go into local Protecting failure state and begin transmission of an SF(1,1) message.
- o A local Manual Switch input SHALL cause the LER to go into local Protecting administrative state and begin transmission of an MS(1,1) message.
- o A local EXER input can be applied when the local end is in remote EXER state. This SHALL cause the LER to remain in the EXER state, but begin transmission of an EXER message instead of RR message.
- o All other local inputs SHALL be ignored.

When in Exercise state, the following describe the reaction to remote messages:

- o A remote Lockout of protection message SHALL cause the LER to go into remote Unavailable state and begin transmission of an NR(0,0) message.
- o A remote Forced Switch message SHALL cause the LER to go into remote Protecting administrative state and begin transmission of an NR(0,1) message.
- o A remote Signal Fail message for the protection path SHALL cause the LER to go into remote Unavailable state and begin transmission of an NR(0,0) message.
- o A remote Signal Fail message for the working path SHALL cause the LER to go into remote Protecting failure state and begin transmission of an NR(0,1) message.
- o A remote Manual Switch message SHALL cause the LER to go into remote Protecting administrative state and begin transmission of an NR(0,1) message.
- o A remote DNR(0,1) message received in remote Exercise state SHALL cause the LER to go into DNR state and begin transmitting DNR(0,1). A remote DNR(0,1) message in local Exercise state is ignored.
- o A remote NR(0,0) message received in remote Exercise state SHALL cause the LER to go into Normal state and begin transmitting NR(0,0). A remote NR message in local Exercise state is ignored.
- o All other local inputs SHALL be ignored.

2.8. Updates to Appendix A. PSC State Machine Tables

Add the following extended states:

E::L = Exercise due to local EXER command
 E::R = Exercise due to remote EXER message

Add the following messages:

State REQ(FP, P)

 E::L EXER(0,0) for revertive, or EXER(0,1) for non-revertive
 E::R RR(0,0) for revertive, or RR(0,1) for non-revertive

Add the following line to the local inputs describing the table description rows:

EXER Exercise

Add the following line to the remote inputs describing the table description rows:

EXER remote Exercise
RR Reverse Request

Modify the state machine as follows (only relevant cells are shown):

Part 1: Local input state machine

	OC	LO	SF-P	FS	SF-W	SFc	MS	WTRE xp	EX ER
N									E: :L
UA:LO :L									i
UA:P: L									i
UA:LO :R									i
UA:P: R									i
PF:W: L									i
PF:W: R									i
PA:F: L									i
PA:M: L									i
PA:F: R									i
PA:M: R									i

WTR									i
DNR									E: :L
E::L	[20]	UA:LO :L	UA:P :L	PA:F :L	PF:W :L	i	PA:M :L	i	i
E::R	i	UA:LO :L	UA:P :L	PA:F :L	PF:W :L	i	PA:M :L	i	E: :L

Part 2: Remote messages state machine

	LO	SF-P	FS	SF-W	MS	WTR	DN R	N R	EX ER	R R
N									E: :R	i
UA:LO :L									i	i
UA:P: L									i	i
UA:LO :R									i	i
UA:P: R									i	i
PF:W: L									i	i
PF:W: R									i	i
PA:F: L									i	i
PA:M: L									i	i
PA:F:									i	i

R										
PA:M: R									i	i
WTR									i	i
DNR									E: :R	i
E::L	UA:LO :R	UA:P :R	PA:F :R	PF:W :R	PA:M :R	i	i	i	i	i
E::R	UA:LO :R	UA:P :R	PA:F :R	PF:W :R	PA:M :R	i	DN R	N	i	i

[20] Transition to N for revertive mode, transition to DNR for non-revertive mode

2.9. Updates to Appendix B. Exercising the Protection Domain

Remove Appendix B.

3. IANA Considerations

This document makes no request of IANA.

Note to RFC Editor: this section may be removed on publication as an RFC.

4. Security Considerations

No specific security issue is raised in addition to those ones already documented in [RFC6378]

5. Acknowledgements

6. References

6.1. Normative References

- [RFC5654] Niven-Jenkins, B., Brungard, D., Betts, M., Sprecher, N., and S. Ueno, "Requirements of an MPLS Transport Profile", RFC 5654, September 2009.

[RFC6378] Weingarten, Y., Bryant, S., Osborne, E., Sprecher, N., and A. Fulignoli, "MPLS Transport Profile (MPLS-TP) Linear Protection", RFC 6378, October 2011.

6.2. Informative References

[LIAISON1205]
ITU-T SG15, , "Liaison Statement: Recommendation ITU-T G.8131/Y.1382 revision - Linear protection switching for MPLS-TP networks ", <https://datatracker.ietf.org/liaison/1205/> , October 2012.

[LIAISON1234]
ITU-T SG15, , "Liaison Statement: Recommendation ITU-T G.8131 revision - Linear protection switching for MPLS-TP networks ", <https://datatracker.ietf.org/liaison/1234/> , February 2013.

Authors' Addresses

Alessandro D'Alessandro
Telecom Italia
via Reiss Romoli, 274
Torino 10141
Italy

Phone: +30 011 2285887
Email: alessandro.dalessandro@telecomitalia.it

Jeong-dong Ryoo
ETRI
218 Gajeongno
Yuseong-gu, Daejeon 305-700
South Korea

Phone: +82-42-860-5384
Email: ryoo@etri.re.kr

Huub van Helvoort
Huawei Technologies
Karspeldreef 4,
Amsterdam 1101 CJ
the Netherlands

Phone: +31 20 4300832
Email: huub.van.helvoort@huawei.com

Internet Engineering Task Force
Internet-Draft
Intended status: Informational
Expires: January 13, 2014

W. George, Ed.
Time Warner Cable
C. Pignataro
R. Asati
K. Raza
Cisco Systems
R. Bonica
Juniper Networks
R. Papneja
D. Dhody
Huawei Technologies
V. Manral
Hewlett-Packard, Inc.
July 12, 2013

Gap Analysis for Operating IPv6-only MPLS Networks
draft-george-mpls-ipv6-only-gap-01

Abstract

This document reviews the MPLS protocol suite in the context of IPv6 and identifies gaps that must be addressed in order to allow MPLS-related protocols and applications to be used with IPv6-only networks. This document is not intended to highlight a particular vendor's implementation (or lack thereof) in the context of IPv6-only MPLS functionality, but rather to focus on gaps in the standards defining the MPLS suite.

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 January 13, 2014.

Copyright Notice

Copyright (c) 2013 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
2. Use Case	4
3. Gap Analysis	5
3.1. MPLS Data Plane	5
3.2. MPLS Control Plane	6
3.2.1. LDP	6
3.2.2. Multicast LDP	6
3.2.3. RSVP- TE	7
3.2.3.1. IGP	7
3.2.3.2. RSVP-TE-P2MP	7
3.2.3.3. RSVP-TE Fast Reroute (FRR)	8
3.2.4. Controller, PCE	8
3.2.5. BGP	8
3.2.6. GMPLS	8
3.3. MPLS Applications	9
3.3.1. L2VPN	9
3.3.1.1. EVPN	9
3.3.2. L3VPN	9
3.3.2.1. 6PE/4PE	10
3.3.2.2. 6VPE/4VPE	10
3.3.2.3. BGP Encapsulation SAFI	10
3.3.2.4. NG-MVPN	11
3.3.3. MPLS-TP	11
3.4. MPLS OAM	11
3.4.1. Extended ICMP	11
3.4.2. LSP Ping	12
3.4.3. BFD	12
3.4.4. Pseudowires	13
3.4.5. MPLS-TP OAM	13
3.5. MIBs	13
4. Gap Summary	13
5. Acknowledgements	14
6. IANA Considerations	14
7. Security Considerations	14
8. Informative References	15
Appendix A. Assignments	20
Authors' Addresses	20

1. Introduction

IPv6 is an integral part of modern network deployments. At the time when this document was written, the majority of these IPv6 deployments were using dual-stack implementations, where IPv4 and IPv6 are supported equally on many or all of the network nodes, and single-stack primarily refers to IPv4-only devices. Dual-stack deployments provide a useful margin for protocols and features that are not currently capable of operating solely over IPv6, because they can continue using IPv4 as necessary. However, as IPv6 deployment and usage becomes more pervasive, and IPv4 exhaustion begins driving changes in address consumption behaviors, there is an increasing likelihood that many networks will need to start operating some or all of their network nodes either as primarily IPv6 (most functions use IPv6, a few legacy features use IPv4), or as IPv6-only (no IPv4 provisioned on the device). This transition toward IPv6-only operation exposes any gaps where features, protocols, or implementations are still reliant on IPv4 for proper function. To that end, and in the spirit of RFC 6540's [RFC6540] recommendation that implementations need to stop requiring IPv4 for proper and complete function, this document reviews the MPLS protocol suite in the context of IPv6 and identifies gaps that must be addressed in order to allow MPLS-related protocols and applications to be used with IPv6-only networks. This document is not intended to highlight a particular vendor's implementation (or lack thereof) in the context of IPv6-only MPLS functionality, but rather to focus on gaps in the standards defining the MPLS suite.

2. Use Case

From a purely theoretical perspective, ensuring that MPLS is fully IP version-agnostic is the right thing to do. However, it is sometimes helpful to understand the underlying drivers that make this work necessary to undertake, especially at a time when IPv6-only networking is still fairly uncommon. This section will discuss some drivers. It is not intended to be a comprehensive discussion of all potential use cases, but rather a discussion of at least one use case so that this is not seen as solving a purely theoretical problem.

IP convergence is continuing to drive new classes of devices to begin communicating via IP. Examples of such devices could include set top boxes for IP Video distribution, cell tower electronics (macro or micro cells), infrastructure Wi-Fi Access Points, and devices for machine to machine (M2M) or Internet of Things applications. In some cases, these classes of devices represent a very large deployment base, on the order of thousands or even millions of devices network-wide. The scale of these networks, coupled with the increasingly

overlapping use of RFC 1918 [RFC1918] address space within the average network, and the lack of globally-routable IPv4 space available for long-term growth begins to drive the need for many of the endpoints in this network to be managed solely via IPv6. Even if these devices are carrying some IPv4 user data, it is often encapsulated in another protocol such that the communication between the endpoint and its upstream devices can be IPv6-only without impacting support for IPv4 on user data. Depending on the MPLS features required, it is plausible to assume that the (existing) MPLS network may need to be extended to these devices.

Additionally, as the impact of IPv4 exhaustion becomes more acute, more and more aggressive IPv4 address reclamation measures will be justified. Measures that were previously seen as too complex or as netting too few addresses for the work required may become more realistic as the cost for obtaining new IPv4 addresses increases. More and more networks are likely to adopt the general stance that IPv4 addresses need to be preserved for revenue-generating customers so that legacy support for IPv4 can be maintained as long as possible. As a result, it may be appropriate for some or all of the network infrastructure, including MPLS LSRs and LERs, to have its IPv4 addresses reclaimed and transition toward IPv6-only operation.

3. Gap Analysis

This gap analysis aims to answer the question, "what breaks when one attempts to use MPLS features on a network of IPv6-only devices?" The assumption is that some endpoints as well as LSRs (PE and P routers) only have IPv6 transport available, and need to support the full suite of MPLS features defined as of the time of this document's writing at parity with the support on an IPv4 network. This is necessary whether they are enabled via LDP RFC 5036 [RFC5036], RSVP-TE RFC 5420 [RFC5420], or BGP RFC 3107 [RFC3107], and whether they are encapsulated in MPLS RFC 3032 [RFC3032], IP RFC 4023 [RFC4023], GRE RFC 4023 [RFC4023], or L2TPv3 RFC 4817 [RFC4817]. It is important when evaluating these gaps to distinguish between user data and control plane data, because while this document is focused on IPv6-only operation, it is quite likely that some amount of the user payload data being carried in the IPv6-only MPLS network will still be IPv4.

3.1. MPLS Data Plane

MPLS labeled packets can be transmitted over a variety of data links RFC 3032 [RFC3032], and MPLS labeled packets can also be encapsulated over IP. The encapsulations of MPLS in IP and Generic Routing Encapsulation (GRE) as well as MPLS over Layer 2 Tunneling Protocol

Version 3 (L2TPv3) support IPv6. See Section 3 of RFC 4023 [RFC4023] and Section 2 of RFC 4817 [RFC4817] respectively.

3.2. MPLS Control Plane

3.2.1. LDP

Label Distribution Protocol (LDP) RFC 5036 [RFC5036] defines a set of procedures for distribution of labels between label switch routers that can use the labels for forwarding traffic. While LDP was designed to use an IPv4 or dual-stack IP network, it has a number of deficiencies that prohibit it from working in an IPv6-only network. LDP-IPv6 [I-D.ietf-mpls-ldp-ipv6] highlights some of the deficiencies when LDP is enabled in IPv6 only or dual-stack networks, and specifies appropriate protocol changes. These deficiencies are related to LSP mapping, LDP identifiers, LDP discovery, LDP session establishment, next hop address and LDP TTL security RFC 5082 [RFC5082].

3.2.2. Multicast LDP

Multipoint LDP (mLDP) is a set of extensions to LDP for setting up Point to Multipoint (P2MP) and Multipoint to Multipoint (MP2MP) LSPs. These extensions are specified in RFC 6388 [RFC6388]. In terms of IPv6-only gap analysis, mLDP has two identified areas of interest:

1. LDP Control plane: Since mLDP uses the LDP control plane to discover and establish sessions with the peer, it shares the same gaps as LDP with regards to control plane (discovery, transport, and session establishment) in an IPv6-only network.
2. Multipoint (MP) FEC Root address: mLDP defines its own MP FECs and rules, different from LDP, to map MP LSPs. mLDP MP FEC contains a Root Address field which is an IP address in IP networks. The current specification allows specifying Root address according to AFI and hence covers both IPv4 or IPv6 root addresses, requiring no extension to support IPv6-only MP LSPs. The root address is used by each LSR participating in an MP LSP setup such that root address reachability is resolved by doing a table lookup against root address to find corresponding upstream neighbor(s). This will pose a problem when an MP LSP traverses islands of IPv4 and IPv4 clouds on the way to the root node.

For example, consider following setup, where R1/R6 are IPv4-only, R3/R4 are IPv6-only, and R2/R5 are dual-stack LSRs:

```
( IPv4-only ) ( IPv6-only ) ( IPv4-only )  
  R1 -- R2 -- R3 -- R4 -- R5 -- R6  
  Leaf                               Root
```

Assume R1 to be a leaf node for an P2MP LSP rooted at R6 (root node). R1 uses R6's IPv4 address as the Root address in MP FEC. As the MP LSP signaling proceeds from R1 to R6, the MP LSP setup will fail on the first IPv6-only transit/branch LSRs (R3) when trying to find IPv4 root address reachability. RFC 6512 [RFC6512] defines a recursive-FEC solution and procedures for mLDP when the backbone (transit/branch) LSRs have no route to the root. The proposed solution is defined for a BGP-free core in an VPN environment, but the similar concept can be used/extended to solve the above issue of IPv6-only backbone receiving an MP FEC element with an IPv4 address. The solution will require a border LSR (the one which is sitting on border of an IPv4/IPv6 island(s) (R2 and R5) to translate an IPv4 root address to equivalent IPv6 address (and vice versa) through the procedures similar to RFC6512. The translation of root address on borders of IPv4 or IPv6 islands will also be needed for recursive FECs and procedures defined in RFC6512.

3.2.3. RSVP- TE

Resource Reservation Protocol Extensions for MPLS Traffic Engineering (RSVP-TE) RFC 3209 [RFC3209] defines a set of procedures & enhancements to establish label-switched tunnels that can be automatically routed away from network failures, congestion, and bottlenecks. RSVP-TE allows establishing an LSP for an IPv4 or IPv6 prefix, thanks to its LSP_TUNNEL_IPv6 object and subobjects.

3.2.3.1. IGP

RFC3630 [RFC3630] specifies a method of adding traffic engineering capabilities to OSPF Version 2. New TLVs and sub-TLVs were added in RFC5329 [RFC5329] to extend TE capabilities to IPv6 networks in OSPF Version 3.

RFC5305 [RFC5305] specifies a method of adding traffic engineering capabilities to IS-IS. New TLVs and sub-TLVs were added in RFC6119 [RFC6119] to extend TE capabilities to IPv6 networks.

3.2.3.2. RSVP-TE-P2MP

RFC4875 [RFC4875] describes extensions to RSVP-TE for the setup of point-to-multipoint (P2MP) LSPs in MPLS and GMPLS with support for both IPv4 and IPv6.

3.2.3.3. RSVP-TE Fast Reroute (FRR)

RFC4090 [RFC4090] specifies FRR mechanisms to establish backup LSP tunnels for local repair supporting both IPv4 and IPv6 networks. Further RFC5286 [RFC5286] describes the use of loop-free alternates to provide local protection for unicast traffic in pure IP and MPLS networks in the event of a single failure, whether link, node, or shared risk link group (SRLG) for both IPv4 and IPv6.

3.2.4. Controller, PCE

The Path Computation Element (PCE) defined in RFC4655 [RFC4655] is an entity that is capable of computing a network path or route based on a network graph, and applying computational constraints. A Path Computation Client (PCC) may make requests to a PCE for paths to be computed. The PCE communication protocol (PCEP) is designed as a communication protocol between PCCs and PCEs for path computations and is defined in RFC5440 [RFC5440].

The PCEP specification RFC5440 [RFC5440] is defined for both IPv4 and IPv6 with support for PCE discovery via an IGP (OSPF RFC5088 [RFC5088], or ISIS RFC5089 [RFC5089]) using both IPv4 and IPv6 addresses. Note that PCEP uses identical encoding of subobjects as in the Resource Reservation Protocol Traffic Engineering Extensions (RSVP-TE) defined in RFC3209 [RFC3209] which supports both IPv4 and IPv6.

The extensions of PCEP to support confidentiality RFC5520 [RFC5520], Route Exclusion RFC5521, [RFC5521] Monitoring RFC5886 [RFC5886], and P2MP RFC6006 [RFC6006] have support for both IPv4 and IPv6.

3.2.5. BGP

RFC3107 [RFC3107] specifies a set of BGP protocol procedures for distributing the labels (for prefixes corresponding to any address-family) between label switch routers so that they can use the labels for forwarding the traffic. RFC3107 allows BGP to distribute the label for IPv4 or IPv6 prefix in an IPv6 only network.

3.2.6. GMPLS

RFC4558 [RFC4558] specifies Node-ID Based RSVP Hello Messages with capability for both IPv4 and IPv6. RFC4990 [RFC4990] clarifies the use of IPv6 addresses in GMPLS networks including handling in the MIB modules.

3.3. MPLS Applications

3.3.1. L2VPN

L2VPN RFC 4664 [RFC4664] specifies two fundamentally different kinds of Layer 2 VPN services that a service provider could offer to a customer: Virtual Private Wire Service (VPWS) and Virtual Private LAN Service (VPLS). RFC 4447 [RFC4447] and RFC 4762 [RFC4762] specify the LDP protocol changes to instantiate VPWS and VPLS services respectively in an MPLS network using LDP as the signaling protocol. This is complemented by RFC 6074 [RFC6074], which specifies a set of procedures for instantiating L2VPNs (e.g. VPWS, VPLS) using BGP as discovery protocol and LDP as well as L2TPv3 as signaling protocol. RFC 4761 [RFC4761] and RFC 6624 [RFC6624] specify BGP protocol changes to instantiate VPLS and VPWS services in an MPLS network, using BGP for both discovery and signaling.

In an IPv6-only MPLS network, use of L2VPN represents connection of Layer 2 islands over an IPv6 MPLS core, and very few changes are necessary to support operation over an IPv6-only network. The L2VPN signaling protocol is either BGP or LDP in an MPLS network, and both can run directly over IPv6 core infrastructure, as well as IPv6 edge devices. RFC 6074 [RFC6074] is the only RFC that appears to have a gap wrt IPv6. In its discovery procedures (section 3.2.2 and section 6), it suggests encoding PE IP address in the VSI-ID, which is encoded in NLRI, which should not exceed 12 bytes (to differentiate its AFI/SAFI encoding from RFC4761). This means that PE IP address can NOT be an IPv6 address. Also, in its signaling procedures (section 3.2.3), it suggests encoding PE_addr in SAII and TAIL, which are limited to 32-bit (AII Type=1) at the moment.

3.3.1.1. EVPN

EVPN [I-D.ietf-l2vpn-evpn] is still a work in progress. As such, it is out of scope for this gap analysis. Instead, the authors of that draft need to ensure that it supports IPv6-only operation, or if it cannot, identify dependencies on underlying gaps in MPLS protocol(s) that must be resolved before it can support IPv6-only operation.

3.3.2. L3VPN

RFC 4364 [RFC4364] defines a method by which a Service Provider may use an IP backbone to provide IP Virtual Private Networks (VPNs) for its customers. The following use cases arise in the context of this gap analysis:

1. Connecting IPv6 islands over IPv6-only MPLS network

2. Connecting IPv4 islands over IPv6-only MPLS network

Both use cases 1 and 2 require mapping an IP packet to an IPv6-signaled LSP to the remote PE, which is not explicitly defined in any RFC. RFC4364 has two MAJOR gaps. First, it is not possible to use an IPv6-only MPLS network, since RFC4364 explicitly assumes IPv4-only MPLS network i.e. BGP Next Hop is assumed to have /32 (for example, see section 5 of RFC4364]. Second, it is limited to VPN-IPv4 address-family i.e. connecting IPv4 islands over IPv4-only MPLS networks. This second gap has been fixed by 6VPE RFC 4659 [RFC4659], which defines connecting IPv6 VPN sites over an IPv4-only MPLS networks, but more work is needed to address the first gap.

The authors do not believe that there are any additional issues encountered when using L2TPv3, RSVP, or GRE (instead of LDP) as transport on an IPv6-only network.

3.3.2.1. 6PE/4PE

RFC 4798 [RFC4798] defines 6PE, which defines how to interconnect IPv6 islands over a Multiprotocol Label Switching (MPLS)-enabled IPv4 cloud. However, use case 2 is doing the opposite, and thus could also be referred to as 4PE. The method to support this use case is not defined explicitly. To support it, IPv4 edge devices need to be able to map IPv4 traffic to MPLS IPv6 core LSP's. Also, the core switches may not understand IPv4 at all, but in some cases they may need to be able to exchange Labeled IPv4 routes from one AS to a neighboring AS.

3.3.2.2. 6VPE/4VPE

RFC 4659 [RFC4659] defines 6VPE, a method by which a Service Provider may use its packet-switched backbone to provide Virtual Private Network (VPN) services for its IPv6 customers. It allows the core network to be MPLS IPv4 or MPLS IPv6, thus addressing use case 1 above. RFC4364 should work as defined for use case 2 above, which could also be referred to as 4VPE, but the RFC does not explicitly discuss this use.

3.3.2.3. BGP Encapsulation SAFI

RFC 5512 [RFC5512] defines the BGP Encapsulation SAFI and the BGP Tunnel Encapsulation Attribute, which can be used to signal tunnelling over an single-Address Family IP core. This mechanism supports transport of MPLS (and other protocols) over Tunnels in an IP core (including an IPv6-only core). In this context, load-balancing can be provided as specified in RFC 5640 [RFC5640].

3.3.2.4. NG-MVPN

TBD RFC 6513 both IPv4 and IPv6 multicast payload traffic

No IP version considerations?

3.3.3. MPLS-TP

***TBD RFC 6371 *** MPLS-TP does not require IP ("and network operation in the absence of a dynamic > control plane or IP forwarding support." RFC 5921) and thus should not be affected by operation on an IPv6-only network.

3.4. MPLS OAM

For MPLS LSPs, there are primarily three OAM mechanisms: Extended ICMP RFC 4884 [RFC4884] RFC 4950 [RFC4950], LSP Ping RFC 4379 [RFC4379], and BFD for MPLS LSPs RFC 5884 [RFC5884]. For MPLS Pseudowires, there is also Virtual Circuit Connectivity Verification (VCCV) RFC 5085 [RFC5085] RFC 5885 [RFC5885]. All of these mechanisms work in pure IPv6 environments. The next subsections cover these in detail.

3.4.1. Extended ICMP

Extended ICMP to support Multi-part messages is defined in RFC 4884 [RFC4884]. This extensibility is defined generally for both ICMPv4 and ICMPv6. The specific ICMP extensions for MPLS are defined in RFC 4950 [RFC4950]. ICMP Multi-part with MPLS extensions works for IPv4 and IPv6. However, the mechanisms described in RFC 4884 and 4950 may fail when tunneling IPv4 traffic over an LSP that is supported by IPv6-only infrastructure.

Assume the following:

- o the path between two IPv4 only hosts contains an MPLS LSP
- o the two routers that terminate the LSP run dual stack
- o the LSP interior routers run IPv6 only
- o the LSP is signaled over IPv6

Now assume that one of the hosts sends an IPv6 packet to the other. However, the packet's TTL expires on an LSP interior router. According to RFC 3032 [RFC3032], the interior router should examine the IPv6 payload, format an ICMPv6 message, and send it (over the tunnel upon which the original packet arrived) to the egress LSP. In

this case, however, the LSP interior router is not IPv6-aware. It cannot parse the original IPv6 datagram, nor can it send an IPv6 message. So, no ICMP message is delivered to the source. Some specific ICMP extensions, in particular ICMP Extensions for Interface and Next-Hop Identification RFC 5837 [RFC5837] restrict the address family of address information included in an Interface Information Object to the same one as the ICMP (see Section 4.5 of RFC 5837). While these extensions are not MPLS specific, they can be used with MPLS packets carrying IP datagrams. This has no implications for IPv6-only environments.

3.4.2. LSP Ping

The LSP Ping mechanism defined in RFC 4379 [RFC4379] is specified to work with IPv6. Specifically, the Target FEC Stacks include both IPv4 and IPv6 versions of all FECs (see Section 3.2 of RFC 4379). The only exceptions are the Pseudowire FECs later specified for IPv6 in RFC 6829 [RFC6829]. Additionally, LSP Ping packets are UDP packets over both IPv4 and IPv6 (see Section 4.3 of RFC 4379). The multipath information includes also IPv6 encodings (see Section 3.3.1 of RFC 4379). However, the mechanisms described in RFC 4379 may fail when tunneling IPv4 traffic over an LSP that is supported by IPv6-only infrastructure.

Assume the following:

- o LSP Ping is operating in traceroute mode over an MPLS LSP
- o the two routers that terminate the LSP run dual stack
- o the LSP interior routers run IPv6 only
- o the LSP is signaled over IPv6

Packets will expire at LSP interior routers. According to RFC 4379, the interior router must parse the IPv4 Echo Request, and then, send an IPv4 Echo Reply. However, the LSP interior router is not IPv4-aware. It cannot parse the IPv4 Echo Request, nor can it send an IPv4 Echo Reply. So, no reply is sent.

3.4.3. BFD

The BFD specification for MPLS LSPs RFC 5884 [RFC5884] is defined for IPv4 as well as IPv6 versions of MPLS FECs (see Section 3.1 of RFC 5884). Additionally the BFD packet is encapsulated over UDP and specified to run over both IPv4 and IPv6 (see Section 7 of RFC 5884).

3.4.4. Pseudowires

The OAM specifications for MPLS Pseudowires define usage for both IPv4 and IPv6. Specifically, VCCV RFC 5085 [RFC5085] can carry IPv4 or IPv6 OAM packets (see Section 5.1.1 and 5.2.1 of RFC 5085), and VCCV for BFD RFC 5885 [RFC5885] also defines an IPv6 encapsulation (see Section 3.2 of RFC 5885).

3.4.5. MPLS-TP OAM

*** TBD***

3.5. MIBs

RFC3811 [RFC3811] defines the textual conventions for MPLS. These lack support for IPv6 in defining MplsExtendedTunnelId and MplsLsrIdentifier. These textual conventions are used in the MPLS TE MIB specification RFC3812 [RFC3812], GMPLS TE MIB specification RFC4802 [RFC4802] and Fast ReRoute (FRR) extension RFC6445 [RFC6445]. 3811bis [I-D.manral-mpls-rfc3811bis] tries to resolve this gap by marking this textual convention as obsolete.

The other MIB specifications for LSR RFC3813 [RFC3813], LDP RFC3815 [RFC3815] and TE RFC4220 [RFC4220] have support for both IPv4 and IPv6.

4. Gap Summary

This draft has reviewed a wide variety of MPLS features and protocols to determine their suitability for use on IPv6-only networks. While some parts of the MPLS suite will function properly without additional changes, gaps have been identified in others, which will need to be addressed with follow-on work. This section will summarize those gaps, along with pointers to any work-in-progress to address them.

Identified gaps in MPLS for IPv6-only networks

Item	Gap	Addressed in
LDP	LSP mapping, LDP identifiers, LDP discovery, LDP session establishment, next hop address and LDP TTL security	LDP-IPv6 [I-D.ietf-mpls-ldp-ipv6]
L2VPN	RFC 6074 [RFC6074] discovery, signaling	TBD
L3VPN	RFC 4364 [RFC4364] BGP next-hop, define method for 4PE/4VPE	TBD
MIBs	RFC 3811 [RFC3811] no IPv6 textual convention	3811bis [I-D.manral-mpls-rfc3811bis]

Table 1: IPv6-only MPLS Gaps

5. Acknowledgements

This draft is brought to you by the letters I, P, V, and the number 6.

6. IANA Considerations

This memo includes no request to IANA.

7. Security Considerations

Changing the address family used for MPLS network operation does not fundamentally alter the security considerations currently extant in any of the specifics of the protocol or its features. However, the change does expose the network and protocol to some of the IPv6-specific security considerations inherent to IPv6 itself as documented in [list of RFCs?]

8. Informative References

- [I-D.ietf-l2vpn-evpn]
Sajassi, A., Aggarwal, R., Henderickx, W., Balus, F., Isaac, A., and J. Uttaro, "BGP MPLS Based Ethernet VPN", draft-ietf-l2vpn-evpn-03 (work in progress), February 2013.
- [I-D.ietf-mpls-ldp-ipv6]
Asati, R., Manral, V., Papneja, R., and C. Pignataro, "Updates to LDP for IPv6", draft-ietf-mpls-ldp-ipv6-08 (work in progress), February 2013.
- [I-D.manral-mpls-rfc3811bis]
Manral, V., Tsou, T., Liu, W., and F. Fondelli, "Definitions of Textual Conventions (TCs) for Multiprotocol Label Switching (MPLS) Management", draft-manral-mpls-rfc3811bis-03 (work in progress), June 2013.
- [RFC1918] Rekhter, Y., Moskowitz, R., Karrenberg, D., Groot, G., and E. Lear, "Address Allocation for Private Internets", BCP 5, RFC 1918, February 1996.
- [RFC3032] Rosen, E., Tappan, D., Fedorkow, G., Rekhter, Y., Farinacci, D., Li, T., and A. Conta, "MPLS Label Stack Encoding", RFC 3032, 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.
- [RFC3630] Katz, D., Kompella, K., and D. Yeung, "Traffic Engineering (TE) Extensions to OSPF Version 2", RFC 3630, September 2003.
- [RFC3811] Nadeau, T. and J. Cucchiara, "Definitions of Textual Conventions (TCs) for Multiprotocol Label Switching (MPLS) Management", RFC 3811, June 2004.
- [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 Router (LSR) Management Information Base (MIB)", RFC 3813, June 2004.
- [RFC3815] Cucchiara, J., Sjostrand, H., and J. Luciani, "Definitions of Managed Objects for the Multiprotocol Label Switching (MPLS), Label Distribution Protocol (LDP)", RFC 3815, June 2004.
- [RFC4023] Worster, T., Rekhter, Y., and E. Rosen, "Encapsulating MPLS in IP or Generic Routing Encapsulation (GRE)", RFC 4023, March 2005.
- [RFC4090] Pan, P., Swallow, G., and A. Atlas, "Fast Reroute Extensions to RSVP-TE for LSP Tunnels", RFC 4090, May 2005.
- [RFC4220] Dubuc, M., Nadeau, T., and J. Lang, "Traffic Engineering Link Management Information Base", RFC 4220, November 2005.
- [RFC4364] Rosen, E. and Y. Rekhter, "BGP/MPLS IP Virtual Private Networks (VPNs)", RFC 4364, February 2006.
- [RFC4379] Kompella, K. and G. Swallow, "Detecting Multi-Protocol Label Switched (MPLS) Data Plane Failures", RFC 4379, February 2006.
- [RFC4447] Martini, L., Rosen, E., El-Aawar, N., Smith, T., and G. Heron, "Pseudowire Setup and Maintenance Using the Label Distribution Protocol (LDP)", RFC 4447, April 2006.
- [RFC4558] Ali, Z., Rahman, R., Prairie, D., and D. Papadimitriou, "Node-ID Based Resource Reservation Protocol (RSVP) Hello: A Clarification Statement", RFC 4558, June 2006.
- [RFC4655] Farrel, A., Vasseur, J., and J. Ash, "A Path Computation Element (PCE)-Based Architecture", RFC 4655, August 2006.
- [RFC4659] De Clercq, J., Ooms, D., Carugi, M., and F. Le Faucheur, "BGP-MPLS IP Virtual Private Network (VPN) Extension for IPv6 VPN", RFC 4659, September 2006.
- [RFC4664] Andersson, L. and E. Rosen, "Framework for Layer 2 Virtual Private Networks (L2VPNs)", RFC 4664, September 2006.
- [RFC4761] Kompella, K. and Y. Rekhter, "Virtual Private LAN Service

- (VPLS) Using BGP for Auto-Discovery and Signaling", RFC 4761, January 2007.
- [RFC4762] Lasserre, M. and V. Kompella, "Virtual Private LAN Service (VPLS) Using Label Distribution Protocol (LDP) Signaling", RFC 4762, January 2007.
- [RFC4798] De Clercq, J., Ooms, D., Prevost, S., and F. Le Faucheur, "Connecting IPv6 Islands over IPv4 MPLS Using IPv6 Provider Edge Routers (6PE)", RFC 4798, February 2007.
- [RFC4802] Nadeau, T. and A. Farrel, "Generalized Multiprotocol Label Switching (GMPLS) Traffic Engineering Management Information Base", RFC 4802, February 2007.
- [RFC4817] Townsley, M., Pignataro, C., Wainner, S., Seely, T., and J. Young, "Encapsulation of MPLS over Layer 2 Tunneling Protocol Version 3", RFC 4817, March 2007.
- [RFC4875] Aggarwal, R., Papadimitriou, D., and S. Yasukawa, "Extensions to Resource Reservation Protocol - Traffic Engineering (RSVP-TE) for Point-to-Multipoint TE Label Switched Paths (LSPs)", RFC 4875, May 2007.
- [RFC4884] Bonica, R., Gan, D., Tappan, D., and C. Pignataro, "Extended ICMP to Support Multi-Part Messages", RFC 4884, April 2007.
- [RFC4950] Bonica, R., Gan, D., Tappan, D., and C. Pignataro, "ICMP Extensions for Multiprotocol Label Switching", RFC 4950, August 2007.
- [RFC4990] Shiimoto, K., Papneja, R., and R. Rabbat, "Use of Addresses in Generalized Multiprotocol Label Switching (GMPLS) Networks", RFC 4990, September 2007.
- [RFC5036] Andersson, L., Minei, I., and B. Thomas, "LDP Specification", RFC 5036, October 2007.
- [RFC5082] Gill, V., Heasley, J., Meyer, D., Savola, P., and C. Pignataro, "The Generalized TTL Security Mechanism (GTSM)", RFC 5082, October 2007.
- [RFC5085] Nadeau, T. and C. Pignataro, "Pseudowire Virtual Circuit Connectivity Verification (VCCV): A Control Channel for Pseudowires", RFC 5085, December 2007.
- [RFC5088] Le Roux, JL., Vasseur, JP., Ikejiri, Y., and R. Zhang,

- "OSPF Protocol Extensions for Path Computation Element (PCE) Discovery", RFC 5088, January 2008.
- [RFC5089] Le Roux, JL., Vasseur, JP., Ikejiri, Y., and R. Zhang, "IS-IS Protocol Extensions for Path Computation Element (PCE) Discovery", RFC 5089, January 2008.
- [RFC5286] Atlas, A. and A. Zinin, "Basic Specification for IP Fast Reroute: Loop-Free Alternates", RFC 5286, September 2008.
- [RFC5305] Li, T. and H. Smit, "IS-IS Extensions for Traffic Engineering", RFC 5305, October 2008.
- [RFC5329] Ishiguro, K., Manral, V., Davey, A., and A. Lindem, "Traffic Engineering Extensions to OSPF Version 3", RFC 5329, September 2008.
- [RFC5420] Farrel, A., Papadimitriou, D., Vasseur, JP., and A. Ayyangarps, "Encoding of Attributes for MPLS LSP Establishment Using Resource Reservation Protocol Traffic Engineering (RSVP-TE)", RFC 5420, February 2009.
- [RFC5440] Vasseur, JP. and JL. Le Roux, "Path Computation Element (PCE) Communication Protocol (PCEP)", RFC 5440, March 2009.
- [RFC5512] Mohapatra, P. and E. Rosen, "The BGP Encapsulation Subsequent Address Family Identifier (SAFI) and the BGP Tunnel Encapsulation Attribute", RFC 5512, April 2009.
- [RFC5520] Bradford, R., Vasseur, JP., and A. Farrel, "Preserving Topology Confidentiality in Inter-Domain Path Computation Using a Path-Key-Based Mechanism", RFC 5520, April 2009.
- [RFC5521] Oki, E., Takeda, T., and A. Farrel, "Extensions to the Path Computation Element Communication Protocol (PCEP) for Route Exclusions", RFC 5521, April 2009.
- [RFC5640] Filss, C., Mohapatra, P., and C. Pignataro, "Load-Balancing for Mesh Softwires", RFC 5640, August 2009.
- [RFC5837] Atlas, A., Bonica, R., Pignataro, C., Shen, N., and JR. Rivers, "Extending ICMP for Interface and Next-Hop Identification", RFC 5837, 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.
- [RFC5886] Vasseur, JP., Le Roux, JL., and Y. Ikejiri, "A Set of Monitoring Tools for Path Computation Element (PCE)-Based Architecture", RFC 5886, June 2010.
- [RFC6006] Zhao, Q., King, D., Verhaeghe, F., Takeda, T., Ali, Z., and J. Meuric, "Extensions to the Path Computation Element Communication Protocol (PCEP) for Point-to-Multipoint Traffic Engineering Label Switched Paths", RFC 6006, September 2010.
- [RFC6074] Rosen, E., Davie, B., Radoaca, V., and W. Luo, "Provisioning, Auto-Discovery, and Signaling in Layer 2 Virtual Private Networks (L2VPNs)", RFC 6074, January 2011.
- [RFC6119] Harrison, J., Berger, J., and M. Bartlett, "IPv6 Traffic Engineering in IS-IS", RFC 6119, February 2011.
- [RFC6388] Wijnands, IJ., Minei, I., Kompella, K., and B. Thomas, "Label Distribution Protocol Extensions for Point-to-Multipoint and Multipoint-to-Multipoint Label Switched Paths", RFC 6388, November 2011.
- [RFC6445] Nadeau, T., Koushik, A., and R. Cetin, "Multiprotocol Label Switching (MPLS) Traffic Engineering Management Information Base for Fast Reroute", RFC 6445, November 2011.
- [RFC6512] Wijnands, IJ., Rosen, E., Napierala, M., and N. Leymann, "Using Multipoint LDP When the Backbone Has No Route to the Root", RFC 6512, February 2012.
- [RFC6540] George, W., Donley, C., Liljenstolpe, C., and L. Howard, "IPv6 Support Required for All IP-Capable Nodes", BCP 177, RFC 6540, April 2012.
- [RFC6624] Kompella, K., Kothari, B., and R. Cherukuri, "Layer 2 Virtual Private Networks Using BGP for Auto-Discovery and Signaling", RFC 6624, May 2012.
- [RFC6829] Chen, M., Pan, P., Pignataro, C., and R. Asati, "Label Switched Path (LSP) Ping for Pseudowire Forwarding Equivalence Classes (FECs) Advertised over IPv6", RFC 6829, January 2013.

Appendix A. Assignments

RFC EDITOR PLEASE REMOVE BEFORE PUBLISHING

This will track which author volunteered for which section(s):

OAM - Ron Bonica, Carlos Pignataro

LDP/mLDP (multicast) - Kamran Raza

L2VPN - Rajiv Asati, Vishwas Manral, Rajiv Papneja

L3VPN - Rajiv Asati, Vishwas Manral, Rajiv Papneja

PCE - Dhruv Dhody, Rajiv Papneja

Editors- Wes George(primary), Vishwas Manral, Rajiv Asati

Authors' Addresses

Wesley George (editor)
Time Warner Cable
13820 Sunrise Valley Drive
Herndon, VA 20111
US

Phone: +1-703-561-2540
Email: wesley.george@twcable.com

Carlos Pignataro
Cisco Systems
7200-12 Kit Creek Road
Research Triangle Park, NC 27709
US

Phone:
Email: cpignata@cisco.com

Rajiv Asati
Cisco Systems
7025 Kit Creek Road
Research Triangle Park, NC 27709
US

Phone:
Email: rajiva@cisco.com

Kamran Raza
Cisco Systems
2000 Innovation Drive
Ottawa, ON K2K-3E8
CA

Phone:
Email: skraza@cisco.com

Ronald Bonica
Juniper Networks
2251 Corporate Park Drive
Herndon, VA 20171
US

Phone:
Email: rbonica@juniper.net

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

Phone:
Email: rajiv.papneja@huawei.com

Dhruv Dhody
Huawei Technologies
2330 Central Expressway
Santa Clara, CA 95050
US

Phone:
Email: dhruv.dhody@huawei.com

Vishwas Manral
Hewlett-Packard, Inc.
19111 Pruneridge Ave.
Cupertino, CA 95014
US

Phone:
Email: vishwas.manral@hp.com

Network Working Group
Internet-Draft
Updates: 3209, 3473 (if approved)
Intended status: Standards Track
Expires: January 16, 2014

K. Kompella
Juniper Networks
M. Hellers
LINX
July 15, 2013

Multi-path Label Switched Paths Signaled Using RSVP-TE
draft-kompella-mpls-rsvp-ecmp-04.txt

Abstract

This document describes extensions to Resource ReSerVation Protocol - Traffic Engineering for the set up of multi-path Traffic Engineered Label Switched Paths (LSPs) in Multi Protocol Label Switching (MPLS) and Generalized MPLS networks, i.e., LSPs that conform to traffic engineering constraints, but follow multiple independent paths from source to destination.

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 January 16, 2014.

Copyright Notice

Copyright (c) 2013 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	2
1.1. Terminology	3
1.2. Conventions used in this document	3
2. Theory of Operation	4
2.1. Multi-path Label Switched Paths	4
2.2. ECMP	5
2.3. Discussion	7
2.4. The Capabilities of TE-based Load Balancing	8
3. Operation of MLSPs	8
3.1. Signaling MLSPs	8
3.2. Label Allocation	8
3.3. Bandwidth Accounting	9
3.4. MLSP Data Plane Actions	10
4. Security Considerations	10
5. Acknowledgments	11
6. IANA Considerations	11
7. References	11
7.1. Normative References	11
7.2. Informative References	11
Authors' Addresses	12

1. Introduction

In selecting a protocol for setting up and signaling "tunnel" Labeled Switched Paths (LSPs) in Multi Protocol Label Switching (MPLS) and Generalized MPLS (GMPLS) networks, one first chooses whether one wants Equal Cost Multi-Path (ECMP) load balancing or Traffic Engineering (TE). For the former, one uses the Label Distribution Protocol (LDP) ([RFC5036]); for the latter, the Resource ReSerVation Protocol - Traffic Engineering (RSVP-TE) ([RFC3209]). [Two other criteria, the need for fast protection and the desire for less configuration, are no longer the deciding factors they used to be, thanks to "IP fast reroute" ([RFC5286]) and "RSVP-TE automesh" ([RFC4972])].

This document describes how one can set up a tunnel LSP that has both ECMP and TE characteristics using RSVP-TE. The techniques described in this document can be used to create a "Multipath LSP" (MLSP) to a destination, that consists of several "sub-LSPs", each potentially taking a different path through the network to the destination. The techniques can also be used to create a single MLSP to multiple equivalent destinations (such as equidistant BGP nexthops announcing a common set of reachable addresses), such that each destination is served by one or more sub-LSPs.

There are several alternatives to choose from when considering MLSPs. One is whether the ingress Label Switching Router (LSR) computes (or otherwise obtains) the full path for each sub-LSP, or whether LSRs along the various paths can compute paths further downstream (using techniques such as "loose hop expansion", as in [RFC5152]). Another is whether the various paths that make up the MLSP have equal cost (or distance) from ingress to egress (i.e., ECMP), whether they may have differing costs. Finally, one can choose whether to terminate a multi-path LSP on a single egress or on several equivalent egresses. For now, the first of each of these alternatives is assumed; future work can explore other choices.

1.1. Terminology

The term Multipath LSP, or MLSP, will be used to denote the (logical) container LSP from an ingress LSR to one or more egress LSR(s). An MLSP is the unit of configuration and management.

An MLSP consists of one or more "sub-LSPs". A sub-LSP consists of a single path from the ingress of the MLS to one of its egresses. A sub-LSP is the unit of signaling of an MLSP. An Explicit Route Object (ERO) will be used to define the path of a sub-LSP.

The "downstream links" of an MLSP Z at LSR X is the union of the downstream links of all sub-LSPs of Z traversing X. Similarly, the "upstream links" of an MLSP Z at LSR X is the union of upstream links of all sub-LSPs of Z traversing X.

The agent that takes the configuration parameters of a tunnel and computes the corresponding paths is called the Path Computation Agent (PCA). The PCA is responsible for acquiring the tunnel configuration, computing the paths of the sub-LSPs, and, if the PCA is not co-located with the ingress, informing the ingress about the tunnel and the EROs for the sub-LSPs.

1.2. 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].

2. Theory of Operation

2.1. Multi-path Label Switched Paths

An MLSP is configured with various constraints associated with TE LSPs, such as destination LSR(s), bandwidth (on a per-class basis, if desired), link colors, Shared Risk Link Groups, etc. [Auto-mesh techniques ([RFC4972]) can be used to reduce configuration; this is not described further here.] In addition, parameters specifically related to MLSPs, such as how many (or the maximum number of) sub-LSPs to create, whether traffic should be split equally across sub-LSPs or not, etc. may also be specified. This configuration lives on the PCA, which is responsible for computing the paths (i.e., the EROs) for the various sub-LSPs. The PCA informs the ingress LSR about the MLSP and the constituent sub-LSPs, including EROs and bandwidths.

The PCA uses the configuration parameters to decide how many sub-LSPs to compute for this MLSP, what paths they should take, and how much bandwidth each sub-LSP is responsible for. Each sub-LSP MUST meet all the constraints of the MLSP (except bandwidth). The bandwidths (per-class, if applicable) of all the sub-LSPs MUST add up to the bandwidth of the MLSP. A Path Computation Element ([RFC4655]) that is multi-path LSP-aware may be used as the PCA.

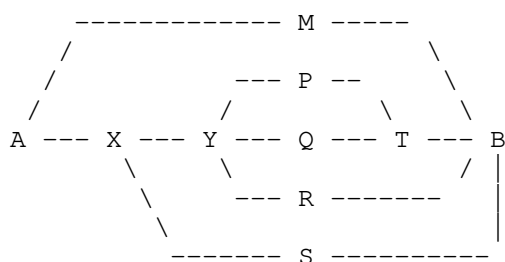
Having computed (or otherwise obtained) the paths of all the sub-LSPs, the ingress A then signals the MLSP by signaling all the individual sub-LSPs across the MPLS/GMPLS network. To do this, the ingress first picks an MLSP ID, a 16-bit number that is unique in the context of the ingress. This ID is used in an ASSOCIATION object that is placed in each sub-LSP to let all transit LSRs know that the sub-LSPs belong to the same MLSP.

If multiple sub-LSPs of the same MLSP pass through LSR Y, and Y has downstream links YP, YQ and YR for the various sub-LSPs, then Y has to load balance incoming traffic for the MLSP across the three downstream links in proportion to the sum of the bandwidths of the sub-LSPs going to each downstream (see Figure 1).

One must distinguish carefully between the signaled bandwidth of a sub-LSP, a static value capturing the expected or maximum traffic on the sub-LSP, and the instantaneous traffic received on a sub-LSP, a constantly varying quantity. Suppose there are three sub-LSPs traversing Y, with bandwidths 10Gbps, 20Gbps and 30Gbps, going to P,

Q and R respectively. Suppose further Y receives some traffic over each of these sub-LSPs. Y must balance this received traffic over the three downstream links YP, YQ and YR in the ratio 1:2:3.

2.2. ECMP



An example network illustrating ECMP. Assume that paths AMB, AXYPB, AXYPB, AXYPB and AXYPB all have the same path length (cost).

Figure 1: Example Network Topology

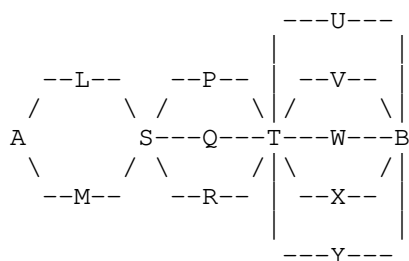
In an IP or LDP network, incoming traffic arriving at A headed for B will be split equally between M and X at A. Similarly, traffic for B arriving at Y will be split equally among P, Q and R. If the traffic arriving at A for B is 120Gbps, then the AMB path will carry 60Gbps, the paths AXYPB, AXYPB and AXYPB will each carry 10Gbps, and the AXYPB path will carry 30Gbps. We'll call this "IP-style" load balancing.

Note: all load balancing is subject to the overriding requirement of mapping the same "flow" to the same downstream. (What constitutes a "flow" is beyond the scope of this document.) This requirement takes precedence over all attempts to balance traffic among downstreams. Thus, the statements above (e.g., "the AMB path will carry 60Gbps") are to be interpreted as ideal targets, not hard requirements, of load balancing.

One can simulate the IP or LDP ECMP behavior with TE-based ECMP by creating an MLSP with five sub-LSPs S1 through S5 taking paths AMB, AXYPB, AXYPB, AXYPB and AXYPB, with bandwidths 60Gbps, 10Gbps, 10Gbps, 10Gbps and 30Gbps, respectively.

With such an arrangement, the MB link carries 60Gbps while the RB link carries just 10Gbps. If one wishes instead to carry equal amounts of traffic on the links incoming to B, then one could arrange the sub-LSPs S1 to S5 to have bandwidths 30Gbps, 15Gbps, 15Gbps, 30Gbps and 30Gbps, respectively. In this case, the bandwidth on each of the four links going to B is 30Gbps, illustrating some of the capabilities of TE-based ECMP.

Staying with this example, A has one sub-LSP of bandwidth 30Gbps to M and four sub-LSPs of total bandwidth 90Gbps to X. Thus, A should load balance traffic in the ratio 1:3 between the AM and the AX links. Similarly, X has three sub-LSPs of total bandwidth 60Gbps to Y and one sub-LSP of bandwidth 30Gbps to S, so X should load balance traffic 2:1 between Y and S. Y has a sub-LSP of bandwidth 15Gbps to each of P and Q and one sub-LSP of bandwidth 30Gbps to R, so Y should load balance traffic 1:1:2 among P, Q and R, respectively. Thus, in general, TE-based ECMP does not assume equal distribution of traffic among downstream LSRs, unlike IP- or LDP-style ECMP.



Another example network illustrating 30 ECMP paths between A and B.

Figure 2: Another Network Topology

In Figure 2, there are potentially $2 \times 3 \times 5 = 30$ ECMP paths between A and B. With IP or LDP, exploiting all these paths is straightforward, and doesn't need a lot of state. With an MLSP as seen so far, this would require 30 sub-LSPs to achieve equivalent load balancing. This suggests that a different approach is needed to efficiently achieve IP-style load balancing with TE LSPs. To this end, we introduce the notion of "equi-bandwidth" (EB) sub-LSPs and EB MLSPs. A sub-LSP is equi-bandwidth if its "E" bit is set (see Section 3.1). An MLSP is equi-bandwidth if all of its sub-LSPs are equi-bandwidth.

If a set of EB sub-LSPs of the same MLSP traverse an LSR S, say to downstream links SP, SQ and SR, then S MUST attempt to load balance traffic received on these EB sub-LSPs equally among the links SP, SQ and SR, independent of how many sub-LSPs go over each of these links. Furthermore, S MUST redistribute traffic received from each of its

upstream LSRs, and SHOULD redistribute all traffic received from upstream as a whole. One can do the former by signaling the same label to each of its upstream LSRs; one can do the latter by signaling the same label to all upstream LSRs (see Section 3.2). For example, in Figure 2, if L sends 12Gbps of traffic to S and M sends 18Gbps to S, S can redistribute L's traffic by sending 4Gbps to each of P, Q and R; and can similarly send 6Gbps of M's traffic to each of P, Q and R. Alternatively, S can load balance the aggregate 30Gbps of traffic received from L and M to each of P, Q and R, thus sending 10Gbps to each. EB sub-LSPs have an added benefit of not requiring unequal load balancing across links, which may pose problems for some hardware.

Given the notion of EB sub-LSPs and EB MLSPs, A can signal an EB MLSP Z comprised of five EB sub-LSPs E1 through E5 with the following paths: ALSPTUB, AMSQTVB, ALSRTWB, AMSPTXB and ALSQTYB (respectively). Then, A has two downstream links for the five sub-LSPs, AL and AM, between which A will load balance equally. Similarly, S has three downstream links, SP, SQ and SR; and T has five downstreams, TU, TV, TW, TX and TY. Thus the load balancing behavior of the MLSP will replicate IP load balancing. The state required for an EB MLSP to achieve IP-style load balancing is somewhat greater than for LDP LSPs, but significantly less than that for multiple "regular" TE LSPs, or for a non-EB MLSP.

2.3. Discussion

Some of the power of TE-based ECMP was illustrated in the above examples. Another is ability to request that all sub-LSPs avoid links colored red. If in the example network in Figure 1, the QT link is colored red but all other links are not, then there are four ECMP paths that satisfy these constraints, and the traffic distribution among them will naturally be different than it would without the link color constraint.

One can also ask whether an MLSP with sub-LSPs is any better than N "regular" LSPs from the same ingress to the same egress. Here are some benefits of an MLSP:

1. With an MLSP, there is a single entity to provision, manage and monitor, versus N separate entities in the case of LSPs. A consequence of this is that with an MLSP, changes in topology can be dealt with easily and autonomously by the ingress LSR, by adding, changing or removing sub-LSPs to rebalance traffic, while maintaining the same TE constraints. With individual LSPs, such changes would require changes in configuration, and thus are harder to automate.

2. An ingress LSR, knowing that an MLSP is for load balancing, can decide on an optimum number of sub-LSPs, and place them appropriately across the network to optimize load balancing. On the other hand, an ingress LSR asked to create N independent LSPs will do so without regard to whether N is a good number of equal cost paths, and, more importantly, may place several of the N LSPs on the same path, defeating the purpose of load balancing.
3. The EB sub-LSP mechanism will, in many cases, result in far fewer sub-LSPs than independent LSPs and thus less control plane state.
4. Finally, an MLSP will usually have less data plane state than N independent LSPs: whenever multiple sub-LSPs traverse a link, a single label will be used for all of them, whereas if multiple LSPs traverse a link, each will need a separate label.

2.4. The Capabilities of TE-based Load Balancing

Definition: Let $G=(V, E)$ be a directed graph (or network), and let A and B in V be two nodes in G. Let T be the traffic arriving at A destined for B. T is said to be "IP-style" load balanced if for every node X on a shortest path from A to B, the portion of T arriving at X is split equally among all nodes Y_i that are adjacent to X and are on a shortest path from X to B.

Theorem: An MLSP can accurately mimic IP-style load balancing between any two nodes in any network.

Proof: left to the reader.

Corollary: MLSPs provide a strictly more powerful load balancing mechanism than IP-style load balancing.

3. Operation of MLSPs

3.1. Signaling MLSPs

Sub-LSPs of an MLSP are tied together using ASSOCIATION objects. ASSOCIATION objects have a new Association Type for MLSPs (TBD). The Association ID is chosen by the ingress of the MLSP; the Association Source is the loopback address of the ingress of the MLSP. All sub-LSPs containing an ASSOCIATION object with a given Association Source and Type belong to the same MLSP.

3.2. Label Allocation

A LSR S that receives Path messages for several sub-LSPs of the same MLSP from the same upstream LSR SHOULD allocate the same label for

all the sub-LSPs. This simplifies load balancing for the aggregate traffic on those sub-LSPs. If the sub-LSPs are EB sub-LSPs, then S SHOULD allocate the same label for all EB sub-LSPs of the same MLSP that pass through S, regardless of which upstream LSR they come from. This allows S to load balance the aggregate traffic received on the MLSP, as all the MLSP traffic arrives at S with the same label. However, an LSR that can achieve the load balancing requirements independent of label allocation strategies is free to do so.

3.3. Bandwidth Accounting

Since MLSPs are traffic engineered, there needs to be strict bandwidth accounting, or admission control, on every link that an MLSP traverses. For non-EB sub-LSPs, this is straightforward, and analogous to regular TE LSPs. However, for EB sub-LSPs, two new procedures are needed, one for signaling bandwidth, and the other for admission control. First, for a given MLSP Z, an LSR X MUST ensure (via signaling) that the total incoming bandwidth of EB sub-LSPs of MLSP Z is divided equally among all the downstream links of X which at least one of the EB sub-LSPs traverses. Second, LSR X MUST ensure that, for each upstream link of X, there is sufficient bandwidth to accommodate all EB sub-LSPs of MLSP Z that traverse that link.

Let's take the example of Figure 2, with MLSP Z having five EB sub-LSPs E1 to E5, and say that MLSP Z is configured with a bandwidth of 30Gbps. Here are some of the steps involved.

1. LSR A, being the ingress, has no upstream links. A has two downstream links, AL and AM. Three EB sub-LSPs of MLSP Z traverse AL, and two traverse AM. A MUST signal a total of 15Gbps for the sub-LSPs to L, and a total of 15Gbps for the sub-LSPs to M. The required bandwidth may be divided up among the sub-LSPs to L (similarly, to M) in any manner so long as the total is 15Gbps. For example, A can signal sub-LSP E1 with 15Gbps, and sub-LSPs E3 and E5 with 0 bandwidth.
2. LSR L has one upstream link AL with three EB sub-LSPs with a total bandwidth of 15Gbps. L MUST ensure that 15Gbps is available for the AL link. If this bandwidth is not available, L MUST send a PathErr on ALL of the EB sub-LSPs on the AL link. Let's assume that the AL link has sufficient bandwidth.
3. Next, it is up to L to decide how to divide the incoming 15Gbps among the three downstream EB sub-LSPs to S. Say L signals sub-LSP E1 with 15Gbps, and the others with 0 bandwidth.
4. LSR S has two upstream links: LS with three EB sub-LSPs with a total bandwidth of 15Gbps, and MS with two EB sub-LSPs with a

total bandwidth of 15Gbps. S MUST ensure that 15Gbps is available for each of the LS and MS links. S has thus a total incoming bandwidth of 30Gbps on MLSP Z. S has to divide this equally among its downstream links SP, SQ and SR, yielding 10Gbps each. S MUST ensure that the total bandwidth requested on the SP link for sub-LSPs E1 and E4 is 10Gbps. S may choose to signal these sub-LSPs with 5Gbps each. Similarly for the SQ and SR links.

There are two important points to note here. One is that the bandwidth reservation (TSpec) for a given EB sub-LSP can (and usually will) change hop-by-hop. The second is that as new EB sub-LSPs are signaled for an MLSP, the bandwidth reservations for existing EB sub-LSPs belonging to the same MLSP may have to be updated. To minimize these updates, it is RECOMMENDED that the first EB sub-LSP on a link be signaled with the total required bandwidth (as far as is known), and later sub-LSPs on the same link be signaled with 0 bandwidth.

3.4. MLSP Data Plane Actions

Traffic intended to be sent over an MLSP is determined at the ingress LSR by means outside the scope of this document, and at transit LSRs by the label(s) assigned by the transit LSR to its upstream LSRs. In the case of non-EB sub-LSPs, this traffic is load balanced across downstream links in the ratio of the bandwidths of the sub-LSPs that comprise the MLSP. In the case of EB sub-LSPs, the traffic belonging to an MLSP from an upstream LSR (or better still, the aggregate traffic for the MLSP from all upstream LSRs) is load balanced equally among all downstream links.

As noted above, the overriding concern is that flows are mapped to the same downstream link (except when the MLSP or some constituent sub-LSPs are changing); this is typically done by hashing fields that define a flow, and mapping hash results to different downstream LSRs. Hash-based load balancing typically assumes that the numbers of flows is sufficiently large and the bandwidth per flow is reasonably well-balanced so that the results of hashing yields reasonable traffic distribution.

Entropy labels ([RFC6790] and [RFC6391]) can be used to improve load balancing at intermediate nodes.

4. Security Considerations

This document introduces no new security concerns in the setup and signaling of LSPs using RSVP-TE, or in the use of the RSVP protocol. [RFC2205] specifies the message integrity mechanisms for RSVP signaling. These mechanisms apply to RSVP-TE signaling of MLSPs

described in this document, and are highly recommended pending newer integrity mechanisms for RSVP.

5. Acknowledgments

The author would like to thank the Routing Protocol group at Juniper Networks for their questions, comments and encouragement for this proposal. While many participated, special thanks go to Yakov Rekhter, John Drake and Rahul Aggarwal. Many thanks too to John for suggesting the use of ASSOCIATION objects.

6. IANA Considerations

IANA is requested to assign a new Association Type for MLSP. This Association Type is to be used for ASSOCIATION objects with C-Type 1 (IPv4 Source) and 2 (IPv6 Source).

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.
- [RFC2205] Braden, B., Zhang, L., Berson, S., Herzog, S., and S. Jamin, "Resource ReSerVation Protocol (RSVP) -- Version 1 Functional Specification", RFC 2205, September 1997.
- [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.
- [RFC4875] Aggarwal, R., Papadimitriou, D., and S. Yasukawa, "Extensions to Resource Reservation Protocol - Traffic Engineering (RSVP-TE) for Point-to-Multipoint TE Label Switched Paths (LSPs)", RFC 4875, May 2007.

7.2. Informative References

- [RFC4655] Farrel, A., Vasseur, J., and J. Ash, "A Path Computation Element (PCE)-Based Architecture", RFC 4655, August 2006.
- [RFC4972] Vasseur, JP., Leroux, JL., Yasukawa, S., Previdi, S., Psenak, P., and P. Mabbey, "Routing Extensions for Discovery of Multiprotocol (MPLS) Label Switch Router (LSR) Traffic Engineering (TE) Mesh Membership", RFC 4972, July 2007.

- [RFC5036] Andersson, L., Minei, I., and B. Thomas, "LDP Specification", RFC 5036, October 2007.
- [RFC5152] Vasseur, JP., Ayyangar, A., and R. Zhang, "A Per-Domain Path Computation Method for Establishing Inter-Domain Traffic Engineering (TE) Label Switched Paths (LSPs)", RFC 5152, February 2008.
- [RFC5226] Narten, T. and H. Alvestrand, "Guidelines for Writing an IANA Considerations Section in RFCs", BCP 26, RFC 5226, May 2008.
- [RFC5286] Atlas, A. and A. Zinin, "Basic Specification for IP Fast Reroute: Loop-Free Alternates", RFC 5286, September 2008.
- [RFC6391] Bryant, S., Filsfils, C., Drafz, U., Kompella, V., Regan, J., and S. Amante, "Flow-Aware Transport of Pseudowires over an MPLS Packet Switched Network", RFC 6391, November 2011.
- [RFC6790] Kompella, K., Drake, J., Amante, S., Henderickx, W., and L. Yong, "The Use of Entropy Labels in MPLS Forwarding", RFC 6790, November 2012.

Authors' Addresses

Kireeti Kompella
Juniper Networks
1194 N. Mathilda Ave.
Sunnyvale, CA 94089
US

Email: kireeti.kompella@gmail.com

Mike Hellers
LINX

Email: mikeh@linx.net

MPLS Working Group
Internet-Draft
Intended status: Standards Track
Expires: January 09, 2014

Chen. Li
Lianyuan. Li
Lu. Huang
China Mobile
Tao. Chou
Quintin. Zhao
Huawei Technology
Emily. Chen

July 08, 2013

Management Information Base for MPLS LDP Multi Topology
draft-li-mpls-ldp-mt-mib-04.txt

Abstract

This memo defines an 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 Multi-Topology Networks over Multi-protocol 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 January 09, 2014.

Copyright Notice

Copyright (c) 2013 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	2
2. The Internet-Standard Management Framework	3
3. Overview of MPLS-LDP-MT-STD-MIB objects	3
3.1. MPLS LDP MT Entity Table	3
3.2. MPLS LDP MT Entity Statistics Table	3
3.3. MPLS LDP MT Session Table	3
3.4. MPLS LDP MT In-segment Tables	4
3.5. MPLS LDP MT Out-segment Tables	4
3.6. MPLS LDP MT LSP Table	4
3.7. MPLS LDP MT Notifications	4
4. MPLS-LDP-MT-STD-MIB Module Definitions	4
5. Security Considerations	27
6. IANA Considerations	27
7. Normative References	27
Authors' Addresses	28

1. Introduction

There are increasing requirements to support multi-topology in MPLS network. For example, service providers want to assign different level of service(s) to different topologies so that the service separation can be achieved. It is also possible to have an in-band management network on top of the original MPLS topology, or maintain separate routing and MPLS domains for isolated multicast or IPv6 islands within the backbone, or force a subset of an address space to follow a different MPLS topology for the purpose of security, QoS or simplified management and/or operations.

For a detailed overview of the multi topology, please refer to I-D .ietf-mpls-ldp-multi-topology.

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 MPLS-LDP-MT-STD-MIB objects

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

3.1. MPLS LDP MT Entity Table

The `mplsLdpEntityTable` specified in [RFC3815] is used to configure information which is used by the LDP protocol to setup potential LDP Sessions. The `mplsLdpMtEntityTable` can be considered as an extension to `mplsLdpEntityTable` to setup potential LDP MT Sessions.

Each entry/row in this table represents a single LDP MT Entity. There is no maximum number of LDP MT Entities specified. However, there is an `mplsLdpMtEntityIndexNext` object which should be retrieved by the command generator prior to creating an LDP MT Entity. If the `mplsLdpMtEntityIndexNext` object is zero, this indicates that the LSR/LER is not able to create another LDP MT Entity at that time.

3.2. MPLS LDP MT Entity Statistics Table

This table provides MPLS Multi Topology performance information on a per-interface basis.

3.3. MPLS LDP MT Session Table

Since all the MT related label messages can be advertised by LDP Sessions in default topology, there is no need to create extra tcp connection for Multi Topology.

The `mplsLdpMtSessionTable` is a read-only table. Each entry in this table represents an MT Session which is related to one or more LDP MT Entities and only one LDP Session in default topology.

3.4. MPLS LDP MT In-segment Tables

The `mplsLdpMtInSegmentTable` contains information about the MPLS Label Distribution Protocol Multi Topology In-Segments which exist on this Label Switching Router (LSR) or Label Edge Router (LER).

The `mplsLdpMtInSegmentStatsTable` contains statistical information for LDP MT in-segments.

3.5. MPLS LDP MT Out-segment Tables

This table contains information about the MPLS Label Distribution Protocol Multi Topology Out-Segments which exist on this Label Switching Router (LSR) or Label Edge Router (LER).

The `mplsLdpMtInSegmentStatsTable` contains statistical information for LDP MT out-segments.

3.6. MPLS LDP MT LSP Table

This table specifies MT LIB label switching information. Entries in this table define LIB label switching entries associated with the specified FEC of the specified topology.

3.7. MPLS LDP MT Notifications

The `mplsLdpMtLspUp` and `mplsLdpMtLspDown` notifications are generated when there is an appropriate change in the `mplsLdpMtLspOperStatus` object, e.g., when the LSP changes state (Up to Down for the `mplsLdpMtLspDown` notification, or Down to Up for the `mplsLdpMtLspUp` notification).

4. MPLS-LDP-MT-STD-MIB Module Definitions

```
MPLS-LDP-MT-STD-MIB DEFINITIONS ::= BEGIN
```

```
IMPORTS
```

```
    IndexIntegerNextFree, IndexInteger
```

```
    FROM DIFFSERV-MIB
```

```
    InetAddress, InetAddressPrefixLength
```

```
    FROM INET-ADDRESS-MIB
```

```
    MplsIndexType
```

```
    FROM MPLS-LSR-STD-MIB
```

```
    MplsLdpLabelType, MplsLspType, MplsLdpIdentifier
```

```
FROM MPLS-TC-STD-MIB
OBJECT-GROUP, MODULE-COMPLIANCE, NOTIFICATION-GROUP
FROM SNMPv2-CONF
transmission, TimeTicks, Integer32, Unsigned32, Counter32,
Counter64, OBJECT-TYPE, MODULE-IDENTITY, NOTIFICATION-TYPE
FROM SNMPv2-SMI
QosService      FROM INTEGRATED-SERVICES-MIB
TimeStamp, StorageType, RowStatus
FROM SNMPv2-TC;

mplsLdpMtStdMIB MODULE-IDENTITY
    LAST-UPDATED "201206131436Z"           -- June 13, 2012 at 14:36 GMT
    ORGANIZATION
        "Multiprotocol Label Switching (mpls) Working Group"
    CONTACT-INFO
        "Chen Li (lichenyj@chinamobile.com)
        Lianyuan Li (lilianyuan@chinamobile.com)
        Lu Huang (huanglu@chinamobile.com)
        China Mobile

        Emily Chen (emily.chenying@huawei.com)
        Quintin Zhao (qzhao@huawei.com)
        Huawei Technologies"
    DESCRIPTION
        "This MIB contains managed object definitions for the
        'Multiprotocol Label Switching, Label Distribution Protocol,
        Multi Topology' document."
    ::= { mplsStdMIB 1 }

--
-- Node definitions
--

-- 1.3.6.1.2.1.10.1.1
    mplsStdMIB OBJECT IDENTIFIER ::= { transmission 166 }

mplsLdpMtNotifications OBJECT IDENTIFIER ::= { mplsLdpMtStdMIB 0 }

mplsLdpMtLspUp NOTIFICATION-TYPE
    OBJECTS { mplsLdpMtLspOperStatus,      -- start of range
              mplsLdpMtLspOperStatus      -- end of range
    }
    ::= { 0 }
```

STATUS current

DESCRIPTION

"This notification is generated when the
mplsLdpMtLspOperStatus object for one or more contiguous
entries in mplsLdpMtLspTable are about to enter the up(1)
state from some other state. The included values of
mplsLdpMtLspOperStatus MUST both be set equal to this new
state (i.e: up(1)). The two instances of
mplsLdpMtLspOperStatus in this notification indicate the rang
e
of indexes that are affected. Note that all the indexes of
the two ends of the range can be derived from the instance
s
identifiers of these two objects. For cases where a contiguous
range of cross-connects have transitioned into the up(1) stat
e
at roughly the same time, the device SHOULD issue a single
notification for each range of contiguous indexes in an effor
t
to minimize the emission of a large number of notifications.
If a notification has to be issued for just a single
cross-connect entry, then the instance identifier (and values
)
of the two mplsLdpMtLspOperStatus objects MUST be the identic
al."

::= { mplsLdpMtNotifications 1 }

mplsLdpMtLspDown NOTIFICATION-TYPE

OBJECTS { mplsLdpMtLspOperStatus, -- start of range
mplsLdpMtLspOperStatus -- end of range
}

STATUS current

DESCRIPTION

"This notification is generated when the
mplsLdpMtLspOperStatus object for one or more contiguous
entries in mplsLdpMtLspTable are about to enter the down(2)
state from some other state. The included values of
mplsLdpMtLspOperStatus MUST both be set equal to this down(2)
state. The two instances of mplsLdpMtLspOperStatus in this
notification indicate the range of indexes that are affected.
Note that all the indexes of the two ends of the range can be
derived from the instance identifiers of these two objects.
For cases where a contiguous range of cross-connects have
transitioned into the down(2) state at roughly the same time,
the device SHOULD issue a single notification for each range
of contiguous indexes in an effort to minimize the emission of
a large number of notifications. If a notification has to be
issued for just a single cross-connect entry, then the
instance identifier (and values) of the two
mplsLdpMtLspOperStatus objects MUST be identical."

::= { mplsLdpMtNotifications 2 }

```
mplsLdpMtObjects OBJECT IDENTIFIER ::= { mplsLdpMtStdMIB 1 }

mplsLdpMtEntityObjects OBJECT IDENTIFIER ::= { mplsLdpMtObjects 1 }

mplsLpMtEntityLastChange OBJECT-TYPE
    SYNTAX TimeStamp
    MAX-ACCESS read-only
    STATUS current
    DESCRIPTION
        "The value of sysUpTime at the time of the most
        recent addition or deletion of an entry
        to/from the mplsLdpMtEntityTable, or
        the most recent change in value of any objects in the
        mplsLdpMtEntityTable.

        If no such changes have occurred since the last
        re-initialization of the local management subsystem,
        then this object contains a zero value."
    ::= { mplsLdpMtEntityObjects 1 }

mplsLdpMtEntityIndexNext OBJECT-TYPE
    SYNTAX IndexIntegerNextFree
    MAX-ACCESS read-only
    STATUS current
    DESCRIPTION
        "This object contains an appropriate value to
        be used for mplsLdpEntityIndex when creating
        entries in the mplsLdpEntityTable. The value
        0 indicates that no unassigned entries are
        available."
    ::= { mplsLdpMtEntityObjects 2 }

-- mplsLdpMtEntityTable
mplsLdpMtEntityTable OBJECT-TYPE
    SYNTAX SEQUENCE OF MplsLdpMtEntityEntry
    MAX-ACCESS read-only
    STATUS current
    DESCRIPTION
        "This table contains information about the
        MPLS Label Distribution Protocol Multi Topology
        Entities which exist on this Label Switching
        Router (LSR) or Label Edge Router (LER)."
    ::= { mplsLdpMtEntityObjects 3 }
```

```
mplsLdpMtEntityEntry OBJECT-TYPE
    SYNTAX MplsLdpMtEntityEntry
    MAX-ACCESS read-only
    STATUS current
    DESCRIPTION
        "An entry in this table represents an LDP MT entity.
        An entry can be created by a network administrator
        or by an SNMP agent as instructed by LDP."
    INDEX { mplsLdpMtEntityLdpId, mplsLdpMtEntityMtId,
            mplsLdpMtEntityIndex }
    ::= { mplsLdpMtEntityTable 1 }

MplsLdpMtEntityEntry ::=
    SEQUENCE {
        mplsLdpMtEntityLdpId
            MplsLdpIdentifier,
        mplsLdpMtEntityMtId
            Unsigned32,
        mplsLdpMtEntityIndex
            IndexInteger,
        mplsLdpMtEntityAdminStatus
            INTEGER,
        mplsLdpMtEntityStorageType
            StorageType,
        mplsLdpMtEntityRowStatus
            RowStatus
    }

mplsLdpMtEntityLdpId OBJECT-TYPE
    SYNTAX MplsLdpIdentifier
    MAX-ACCESS read-only
    STATUS current
    DESCRIPTION
        "The LDP identifier."
    REFERENCE
        "RFC 5036, LDP Specification, Section on LDP Identifiers."
    ::= { mplsLdpMtEntityEntry 1 }

mplsLdpMtEntityMtId OBJECT-TYPE
    SYNTAX Unsigned32 (0..65535)
    MAX-ACCESS read-only
    STATUS current
    DESCRIPTION
        "The Multi Topology identifier of this LDP MT Entity."
    REFERENCE
        "draft-ietf-mpls-ldp-multi-topology, LDP Extensions for Multi
```

Topology Routing, Section on Multi-Topology ID."
 ::= { mplsLdpMtEntityEntry 2 }

mplsLdpMtEntityIndex OBJECT-TYPE

SYNTAX IndexInteger

MAX-ACCESS read-only

STATUS current

DESCRIPTION

"This index is used as a secondary index to uniquely identify this row. Before creating a row in this table, the 'mplsLdpMtEntityIndexNext' object should be retrieved. That value should be used for the value of this index when creatin

g

a row in this table. NOTE: if a value of zero (0) is retrieved, that indicates that no rows can be created in this table at this time."

::= { mplsLdpMtEntityEntry 3 }

mplsLdpMtEntityAdminStatus OBJECT-TYPE

SYNTAX INTEGER

{
 enable(1),
 disable(2)
}

MAX-ACCESS read-only

STATUS current

DESCRIPTION

"The administrative status of this LDP MT Entity. If this object is changed from 'enable' to 'disable' and this entity has already attempted to establish contact with a MT Session, then all contact with that MT Session is lost and all information from that MT Session needs to be removed from the MIB. (This implies that the network management subsystem should clean up any related entry in the mplsLdpMtSessionTable.). At this point the operator is able to change values which are related to this entity. When the admin status is set back to 'enable', then this MT Entity wil

1

attempt to establish a new MT Session."

DEFVAL { enable }

::= { mplsLdpMtEntityEntry 4 }

mplsLdpMtEntityStorageType OBJECT-TYPE

SYNTAX StorageType

MAX-ACCESS read-only

STATUS current

DESCRIPTION

"The storage type for this conceptual row. Conceptual rows having the value 'permanent(4)' need not allow write-access to any columnar objects in the row."

::= { mplsLdpMtEntityEntry 5 }

mplsLdpMtEntityRowStatus OBJECT-TYPE

SYNTAX RowStatus

MAX-ACCESS read-only

STATUS current

DESCRIPTION

"The status of this conceptual row. All writable objects in this row may be modified at any time, however, as described in detail in the section entitled, 'Changing Values After Session Establishment', and again described in the DESCRIPTION

N

clause of the mplsLdpMtEntityAdminStatus object, if a session has been initiated with a Peer, changing objects in this tabl

e

will wreak havoc with the session and interrupt traffic. To repeat again: the recommended procedure is to set the mplsLdpMtEntityAdminStatus to down, thereby explicitly causin

g

a session to be torn down. Then, change objects in this entr

y,

then set the mplsLdpMtEntityAdminStatus to enable, which enab

les

a new session to be initiated."

::= { mplsLdpMtEntityEntry 6 }

-- mplsLdpMtEntityStatsTable

mplsLdpMtEntityStatsTable OBJECT-TYPE

SYNTAX SEQUENCE OF MplsLdpMtEntityStatsEntry

MAX-ACCESS read-only

STATUS current

DESCRIPTION

"This table contains statistical information for LDP MT entities to an LSR."

::= { mplsLdpMtEntityObjects 4 }

mplsLdpMtEntityStatsEntry OBJECT-TYPE

SYNTAX MplsLdpMtEntityStatsEntry

MAX-ACCESS read-only

STATUS current

DESCRIPTION

"An entry in this table is created by the LSR for every interface capable of supporting MPLS LDP Multi Topology. It is an extension to the mplsLdpMtEntityEntry table. Note that the discontinuity behavior of entries in this table MUST be based on the corresponding ifEntry's ifDiscontinuityTime."

```
AUGMENTS { mplsLdpMtEntityEntry }  
::= { mplsLdpMtEntityStatsTable 1 }
```

```
MplsLdpMtEntityStatsEntry ::=   
    SEQUENCE {  
        mplsLdpMtEntityStatsOctets  
            Counter32,  
        mplsLdpMtEntityStatsPackets  
            Counter32,  
        mplsLdpMtEntityStatsErrors  
            Counter32,  
        mplsLdpMtEntityStatsDiscards  
            Counter32,  
        mplsLdpMtEntityStatsHCOctets  
            Counter64,  
        mplsLdpMtEntityStatsDiscontinuityTime  
            TimeTicks  
    }
```

```
mplsLdpMtEntityStatsOctets OBJECT-TYPE  
    SYNTAX Counter32  
    MAX-ACCESS read-only  
    STATUS current  
    DESCRIPTION  
        "This value represents the total number of octets received  
        by this MT interface. It MUST be equal to the least  
        significant 32 bits of mplsLdpMtEntityStatsHCOctets if  
        mplsLdpMtEntityStatsHCOctets is supported according to  
        the rules spelled out in RFC2863."  
    ::= { mplsLdpMtEntityStatsEntry 1 }
```

```
mplsLdpMtEntityStatsPackets OBJECT-TYPE  
    SYNTAX Counter32  
    MAX-ACCESS read-only  
    STATUS current  
    DESCRIPTION  
        "Total number of packets received by this MT interface."  
    ::= { mplsLdpMtEntityStatsEntry 2 }
```

```
mplsLdpMtEntityStatsErrors OBJECT-TYPE  
    SYNTAX Counter32  
    MAX-ACCESS read-only  
    STATUS current  
    DESCRIPTION  
        "The number of error packets received on this MT interface."
```

```
::= { mplsLdpMtEntityStatsEntry 3 }
```

```
mplsLdpMtEntityStatsDiscards OBJECT-TYPE
```

```
SYNTAX Counter32
```

```
MAX-ACCESS read-only
```

```
STATUS current
```

```
DESCRIPTION
```

"The number of labeled packets received on this MT interface, which were chosen to be discarded even though no errors had been detected to prevent their being transmitted.

One possible reason for discarding such a labeled packet could be to free up buffer space."

```
::= { mplsLdpMtEntityStatsEntry 4 }
```

```
mplsLdpMtEntityStatsHCOctets OBJECT-TYPE
```

```
SYNTAX Counter64
```

```
MAX-ACCESS read-only
```

```
STATUS current
```

```
DESCRIPTION
```

"The total number of octets received. This is the 64 bit version of mplsLdpMtEntityStatsOctets, if

mplsLdpMtEntityStatsHCOctets is supported according to the rules spelled out in RFC2863."

```
::= { mplsLdpMtEntityStatsEntry 5 }
```

```
mplsLdpMtEntityStatsDiscontinuityTime OBJECT-TYPE
```

```
SYNTAX TimeTicks
```

```
MAX-ACCESS read-only
```

```
STATUS current
```

```
DESCRIPTION
```

"The value of sysUpTime on the most recent occasion at which any one or more of this MT interface's Counter32 or Counter64 suffered a discontinuity. If no such discontinuities have occurred since the last re-initialization of the local management subsystem, then this object contains a zero value.

"

```
::= { mplsLdpMtEntityStatsEntry 6 }
```

```
mplsLdpMtSessionObjects OBJECT IDENTIFIER
```

```
::= { mplsLdpMtObjects 2 }
```

```
mplsLdpMtSessionLastChange OBJECT-TYPE
```

```
SYNTAX TimeStamp
```

```
MAX-ACCESS read-only
```

```

STATUS current
DESCRIPTION
    "The value of sysUpTime at the time of the most
    recent addition or deletion to/from the
    mplsLdpMtSessionTable."
::= { mplsLdpMtSessionObjects 1 }

-- mplsLdpMtSessionTable
mplsLdpMtSessionTable OBJECT-TYPE
    SYNTAX SEQUENCE OF MplsLdpMtSessionEntry
    MAX-ACCESS read-only
    STATUS current
    DESCRIPTION
        "A table of MT Sessions between the LDP MT Entities.  Each ro
w
        in this table represents a single MT session."
    ::= { mplsLdpMtSessionObjects 2 }

mplsLdpMtSessionEntry OBJECT-TYPE
    SYNTAX MplsLdpMtSessionEntry
    MAX-ACCESS read-only
    STATUS current
    DESCRIPTION
        "An entry in this table represents information on a single MT
        session.  The information contained in a row is read-only."
    INDEX { mplsLdpMtEntityLdpId, mplsLdpMtEntityMtId,
            mplsLdpMtEntityIndex, mplsLdpMtSessionPeerId }
    ::= { mplsLdpMtSessionTable 1 }

MplsLdpMtSessionEntry ::=
    SEQUENCE {
        mplsLdpMtSessionPeerId
            MplsLdpIdentifier,
        mplsLdpMtSessionState
            INTEGER,
        mplsLdpMtSessionStateLastChange
            TimeStamp
    }

mplsLdpMtSessionPeerId OBJECT-TYPE
    SYNTAX MplsLdpIdentifier
    MAX-ACCESS read-only
    STATUS current
    DESCRIPTION
        "The LDP identifier of this LDP MT Peer."
    ::= { mplsLdpMtSessionEntry 1 }

```

```

mplsLdpMtSessionState OBJECT-TYPE
    SYNTAX INTEGER
        {
            initialized(1),
            operational(2)
        }
    MAX-ACCESS read-only
    STATUS current
    DESCRIPTION
        "The current state of the MT Session.  When the tcp connectio
n
        in default topology is established, and both ends have the
        capability of the given MT-ID, the state can change from
        initialized to operational."
    ::= { mplsLdpMtSessionEntry 2 }

mplsLdpMtSessionStateLastChange OBJECT-TYPE
    SYNTAX TimeStamp
    MAX-ACCESS read-only
    STATUS current
    DESCRIPTION
        "The value of sysUpTime at the time this MT Session was
        created."
    ::= { mplsLdpMtSessionEntry 3 }

mplsLdpMtLspObjects OBJECT IDENTIFIER ::= { mplsLdpMtObjects 3 }

-- mplsLdpMtInSegmentTable
mplsLdpMtInSegmentTable OBJECT-TYPE
    SYNTAX SEQUENCE OF MplsLdpMtInSegmentEntry
    MAX-ACCESS read-only
    STATUS current
    DESCRIPTION
        "This table contains information about the MPLS Label
        Distribution Protocol Multi Topology
        In-Segments which exist on this Label Switching Router (LSR)
        or Label Edge Router (LER)."
    ::= { mplsLdpMtLspObjects 1 }

mplsLdpMtInSegmentEntry OBJECT-TYPE
    SYNTAX MplsLdpMtInSegmentEntry
    MAX-ACCESS read-only
    STATUS current
    DESCRIPTION
        "An entry in this table represents information on a single

```

LDP MT LSP which is represented by a MT session's index combination (mplsLdpMtEntityLdpId, mplsLdpMtEntityMtId, mplsLdpMtEntityIndex, mplsLdpMtSessionPeerId).

The information contained in a row is read-only."
INDEX { mplsLdpMtEntityLdpId, mplsLdpMtEntityMtId,
mplsLdpMtEntityIndex, mplsLdpMtSessionPeerId }
::= { mplsLdpMtInSegmentTable 1 }

MplsLdpMtInSegmentEntry ::=

```
SEQUENCE {  
    mplsLdpMtInSegmentIndex  
        MplsIndexType,  
    mplsLdpMtInSegmentLabelType  
        MplsLdpLabelType,  
    mplsLdpMtInSegmentLspType  
        MplsLspType  
}
```

mplsLdpMtInSegmentIndex OBJECT-TYPE

SYNTAX MplsIndexType

MAX-ACCESS read-only

STATUS current

DESCRIPTION

"The index for this MT in-segment. The string containing the single octet 0x00 MUST not be used as an index."

::= { mplsLdpMtInSegmentEntry 1 }

mplsLdpMtInSegmentLabelType OBJECT-TYPE

SYNTAX MplsLdpLabelType

MAX-ACCESS read-only

STATUS current

DESCRIPTION

"The Layer 2 Label Type."

::= { mplsLdpMtInSegmentEntry 2 }

mplsLdpMtInSegmentLspType OBJECT-TYPE

SYNTAX MplsLspType

MAX-ACCESS read-only

STATUS current

DESCRIPTION

"The type of LSP connection."

::= { mplsLdpMtInSegmentEntry 3 }

```

-- mplsLdpMtInSegmentStatsTable
    mplsLdpMtInSegmentStatsTable OBJECT-TYPE
        SYNTAX SEQUENCE OF MplsLdpMtInSegmentStatsEntry
        MAX-ACCESS read-only
        STATUS current
        DESCRIPTION
            "This table contains statistical information for LDP MT
            in-segments to an LSR."
        ::= { mplsLdpMtLspObjects 2 }

mplsLdpMtInSegmentStatsEntry OBJECT-TYPE
    SYNTAX MplsLdpMtInSegmentStatsEntry
    MAX-ACCESS read-only
    STATUS current
    DESCRIPTION
        "An entry in this table contains statistical information about
        one incoming MT segment which is configured in the
        mplsLdpMtInSegmentTable. The counters in this entry should
        behave in a manner similar to that of the MT interface.
        mplsLdpMtInSegmentStatsDiscontinuityTime indicates the time
        of the last discontinuity in all of these objects."
    AUGMENTS { mplsLdpMtInSegmentEntry }
    ::= { mplsLdpMtInSegmentStatsTable 1 }

MplsLdpMtInSegmentStatsEntry ::=
    SEQUENCE {
        mplsLdpMtInSegmentStatsOctets
            Counter32,
        mplsLdpMtInSegmentStatsPackets
            Counter32,
        mplsLdpMtInSegmentStatsErrors
            Counter32,
        mplsLdpMtInSegmentStatsDiscards
            Counter32,
        mplsLdpMtInSegmentStatsHCOctets
            Counter64,
        mplsLdpMtInSegmentStatsDiscontinuityTime
            TimeTicks
    }

mplsLdpMtInSegmentStatsOctets OBJECT-TYPE
    SYNTAX Counter32
    MAX-ACCESS read-only
    STATUS current
    DESCRIPTION
        "This value represents the total number of octets received

```

by this MT segment. It MUST be equal to the least significant 32 bits of mplsLdpMtInSegmentStatsHCOctets if mplsLdpMtInSegmentStatsHCOctets is supported according to the rules spelled out in RFC2863."
 ::= { mplsLdpMtInSegmentStatsEntry 1 }

mplsLdpMtInSegmentStatsPackets OBJECT-TYPE
 SYNTAX Counter32
 MAX-ACCESS read-only
 STATUS current
 DESCRIPTION
 "Total number of packets received by this MT segment."
 ::= { mplsLdpMtInSegmentStatsEntry 2 }

mplsLdpMtInSegmentStatsErrors OBJECT-TYPE
 SYNTAX Counter32
 MAX-ACCESS read-only
 STATUS current
 DESCRIPTION
 "The number of error packets received on this MT segment."
 ::= { mplsLdpMtInSegmentStatsEntry 3 }

mplsLdpMtInSegmentStatsDiscards OBJECT-TYPE
 SYNTAX Counter32
 MAX-ACCESS read-only
 STATUS current
 DESCRIPTION
 "The number of labeled packets received on this MT in-segment
 ,
 which were chosen to be discarded even though no errors had
 been detected to prevent their being transmitted.
 One possible reason for discarding such a labeled packet
 could be to free up buffer space."
 ::= { mplsLdpMtInSegmentStatsEntry 4 }

mplsLdpMtInSegmentStatsHCOctets OBJECT-TYPE
 SYNTAX Counter64
 MAX-ACCESS read-only
 STATUS current
 DESCRIPTION
 "The total number of octets received. This is the 64 bit
 version of mplsLdpMtInSegmentStatsOctets, if
 mplsLdpMtInSegmentStatsHCOctets is supported according to the
 rules spelled out in RFC2863."
 ::= { mplsLdpMtInSegmentStatsEntry 5 }

```

mplsLdpMtInSegmentStatsDiscontinuityTime OBJECT-TYPE
    SYNTAX TimeTicks
    MAX-ACCESS read-only
    STATUS current
    DESCRIPTION
        "The value of sysUpTime on the most recent occasion at which
        any one or more of this MT segment's Counter32 or Counter64
        suffered a discontinuity. If no such discontinuities have
        occurred since the last re-initialization of the local
        management subsystem, then this object contains a zero value."
    ::= { mplsLdpMtInSegmentStatsEntry 6 }

-- mplsLdpMtOutSegmentTable
mplsLdpMtOutSegmentTable OBJECT-TYPE
    SYNTAX SEQUENCE OF MplsLdpMtOutSegmentEntry
    MAX-ACCESS read-only
    STATUS current
    DESCRIPTION
        "This table contains information about the MPLS Label
        Distribution Protocol Multi Topology Out-Segments which
        exist on this Label Switching Router (LSR) or Label
        Edge Router (LER)."
    ::= { mplsLdpMtLspObjects 3 }

mplsLdpMtOutSegmentEntry OBJECT-TYPE
    SYNTAX MplsLdpMtOutSegmentEntry
    MAX-ACCESS read-only
    STATUS current
    DESCRIPTION
        "An entry in this table represents information on a single
        LDP MT LSP which is represented by a MT session's index
        combination (mplsLdpMtEntityLdpId, mplsLdpMtEntityMtId,
        mplsLdpMtEntityIndex, mplsLdpMtSessionPeerId).

        The information contained in a row is read-only."
    INDEX { mplsLdpMtEntityLdpId, mplsLdpMtEntityMtId,
            mplsLdpMtEntityIndex, mplsLdpMtSessionPeerId }
    ::= { mplsLdpMtOutSegmentTable 1 }

MplsLdpMtOutSegmentEntry ::=
    SEQUENCE {
        mplsLdpMtOutSegmentIndex
            MplsIndexType,
        mplsLdpMtOutSegmentLabelType
            MplsLdpLabelType,

```

```
        mplsLdpMtOutSegmentLspType
            MplsLspType
    }

mplsLdpMtOutSegmentIndex OBJECT-TYPE
    SYNTAX MplsIndexType
    MAX-ACCESS read-only
    STATUS current
    DESCRIPTION
        "The index for this MT out-segment. The string containing
        the single octet 0x00 MUST not be used as an index."
    ::= { mplsLdpMtOutSegmentEntry 1 }

mplsLdpMtOutSegmentLabelType OBJECT-TYPE
    SYNTAX MplsLdpLabelType
    MAX-ACCESS read-only
    STATUS current
    DESCRIPTION
        "The Layer 2 Label Type."
    ::= { mplsLdpMtOutSegmentEntry 2 }

mplsLdpMtOutSegmentLspType OBJECT-TYPE
    SYNTAX MplsLspType
    MAX-ACCESS read-only
    STATUS current
    DESCRIPTION
        "The type of LSP connection."
    ::= { mplsLdpMtOutSegmentEntry 3 }

-- mplsLdpMtOutSegmentStatsTable
mplsLdpMtOutSegmentStatsTable OBJECT-TYPE
    SYNTAX SEQUENCE OF MplsLdpMtOutSegmentStatsEntry
    MAX-ACCESS read-only
    STATUS current
    DESCRIPTION
        "This table contains statistical information for LDP MT
        out-segments to an LSR."
    ::= { mplsLdpMtLspObjects 4 }

mplsLdpMtOutSegmentStatsEntry OBJECT-TYPE
    SYNTAX MplsLdpMtOutSegmentStatsEntry
    MAX-ACCESS read-only
    STATUS current
    DESCRIPTION
```

"An entry in this table contains statistical information about one incoming MT segment which is configured in the mplsLdpMtOutSegmentTable. The counters in this entry should behave in a manner similar to that of the MT interface. mplsLdpMtOutSegmentStatsDiscontinuityTime indicates the time of the last discontinuity in all of these objects."

AUGMENTS { mplsLdpMtOutSegmentEntry }
::= { mplsLdpMtOutSegmentStatsTable 1 }

MplsLdpMtOutSegmentStatsEntry ::=

SEQUENCE {
 mplsLdpMtOutSegmentStatsOctets
 Counter32,
 mplsLdpMtOutSegmentStatsPackets
 Counter32,
 mplsLdpMtOutSegmentStatsErrors
 Counter32,
 mplsLdpMtOutSegmentStatsDiscards
 Counter32,
 mplsLdpMtOutSegmentStatsHCOctets
 Counter64,
 mplsLdpMtOutSegmentStatsDiscontinuityTime
 TimeTicks
}

mplsLdpMtOutSegmentStatsOctets OBJECT-TYPE
SYNTAX Counter32
MAX-ACCESS read-only
STATUS current
DESCRIPTION
 "This value represents the total number of octets received by this MT segment. It MUST be equal to the least significant 32 bits of mplsLdpMtOutSegmentStatsHCOctets if mplsLdpMtOutSegmentStatsHCOctets is supported according to the rules spelled out in RFC2863."
::= { mplsLdpMtOutSegmentStatsEntry 1 }

mplsLdpMtOutSegmentStatsPackets OBJECT-TYPE
SYNTAX Counter32
MAX-ACCESS read-only
STATUS current
DESCRIPTION
 "Total number of packets received by this MT segment."
::= { mplsLdpMtOutSegmentStatsEntry 2 }

```

mplsLdpMtOutSegmentStatsErrors OBJECT-TYPE
    SYNTAX Counter32
    MAX-ACCESS read-only
    STATUS current
    DESCRIPTION
        "The number of error packets received on this MT segment."
    ::= { mplsLdpMtOutSegmentStatsEntry 3 }

mplsLdpMtOutSegmentStatsDiscards OBJECT-TYPE
    SYNTAX Counter32
    MAX-ACCESS read-only
    STATUS current
    DESCRIPTION
        "The number of labeled packets received on this MT out-segmen
t,
        which were chosen to be discarded even though no errors had
        been detected to prevent their being transmitted.
        One possible reason for discarding such a labeled packet
        could be to free up buffer space."
    ::= { mplsLdpMtOutSegmentStatsEntry 4 }

mplsLdpMtOutSegmentStatsHCOctets OBJECT-TYPE
    SYNTAX Counter64
    MAX-ACCESS read-only
    STATUS current
    DESCRIPTION
        "The total number of octets received. This is the 64 bit
        version of mplsLdpMtOutSegmentStatsOctets, if
        mplsLdpMtOutSegmentStatsHCOctets is supported according to
        the rules spelled out in RFC2863."
    ::= { mplsLdpMtOutSegmentStatsEntry 5 }

mplsLdpMtOutSegmentStatsDiscontinuityTime OBJECT-TYPE
    SYNTAX TimeTicks
    MAX-ACCESS read-only
    STATUS current
    DESCRIPTION
        "The value of sysUpTime on the most recent occasion at which
        any one or more of this MT segment's Counter32 or Counter64
        suffered a discontinuity. If no such discontinuities have
        occurred since the last re-initialization of the local
        management subsystem, then this object contains a zero value."
    ::= { mplsLdpMtOutSegmentStatsEntry 6 }

mplsLdpMtLspLastChange OBJECT-TYPE

```

```
SYNTAX TimeStamp
MAX-ACCESS read-only
STATUS current
DESCRIPTION
    "The value of sysUpTime at the time of the most recent additi
on
    or deletion of an entry to/from the mplsLdpMtLspTable, or the
    most recent change in value of any objects in the
    mplsLdpMtLspTable.

    If no such changes have occurred since the last
    re-initialization of the local management subsystem,
    then this object contains a zero value."
 ::= { mplsLdpMtLspObjects 5 }

mplsLdpMtLspIndexNext OBJECT-TYPE
    SYNTAX IndexIntegerNextFree
    MAX-ACCESS read-only
    STATUS current
    DESCRIPTION
        "This object contains an appropriate value to be used for
        mplsLdpMtLspIndex when creating entries in the
        mplsLdpMtLspTable. The value 0 indicates that no unassigned
        entries are available."
    ::= { mplsLdpMtLspObjects 6 }

-- mplsLdpMtLspTable
mplsLdpMtLspTable OBJECT-TYPE
    SYNTAX SEQUENCE OF MplsLdpMtLspEntry
    MAX-ACCESS read-only
    STATUS current
    DESCRIPTION
        "This table specifies MT LIB label switching information.
        Entries in this table define LIB label switching entries
        associated with the specified topology."
    ::= { mplsLdpMtLspObjects 7 }

mplsLdpMtLspEntry OBJECT-TYPE
    SYNTAX MplsLdpMtLspEntry
    MAX-ACCESS read-only
    STATUS current
    DESCRIPTION
        "An entry in this table is created by an LSR for every label
        within the context of a specific topology capable of
        supporting MT LDP LSP. The indexing provides an ordering
        of topologies per interface."
```

```
INDEX { mplsLdpMtEntityLdpId, mplsLdpMtEntityMtId,  
mplsLdpMtEntityIndex, mplsLdpMtLspInSegmentIndex,  
mplsLdpMtLspOutSegmentIndex, mplsLdpMtLspIndex }  
::= { mplsLdpMtLspTable 1 }
```

```
MplsLdpMtLspEntry ::=   
SEQUENCE {  
    mplsLdpMtLspIndex  
        IndexInteger,  
    mplsLdpMtLspFecAddr  
        InetAddress,  
    mplsLdpMtLspFecAddrLength  
        InetAddressPrefixLength,  
    mplsLdpMtLspInSegmentIndex  
        MplsIndexType,  
    mplsLdpMtLspOutSegmentIndex  
        MplsIndexType,  
    mplsLdpMtLspRowStatus  
        Integer32,  
    mplsLdpMtLspStorageType  
        StorageType,  
    mplsLdpMtLspOperStatus  
        RowStatus,  
    mplsLdpMtLspService  
        QosService  
}
```

```
mplsLdpMtLspIndex OBJECT-TYPE  
SYNTAX IndexInteger  
MAX-ACCESS read-only  
STATUS current  
DESCRIPTION  
    "The index which uniquely identifies this entry."  
::= { mplsLdpMtLspEntry 1 }
```

```
mplsLdpMtLspFecAddr OBJECT-TYPE  
SYNTAX InetAddress  
MAX-ACCESS read-only  
STATUS current  
DESCRIPTION  
    "The FEC address of this LDP MT LSP. Note that the  
    value of this object is interpreted as prefix address."  
REFERENCE  
    "RFC 5036, Section 3.4.1 FEC TLV."  
::= { mplsLdpMtLspEntry 2 }
```

```
mplsLdpMtLspFecAddrLength OBJECT-TYPE
    SYNTAX InetAddressPrefixLength
    MAX-ACCESS read-only
    STATUS current
    DESCRIPTION
        "The FEC prefix length of this LDP MT LSP."
    REFERENCE
        "RFC5036, Section 3.4.1. FEC TLV"
    ::= { mplsLdpMtLspEntry 3 }

mplsLdpMtLspInSegmentIndex OBJECT-TYPE
    SYNTAX MplsIndexType
    MAX-ACCESS read-only
    STATUS current
    DESCRIPTION
        "Index of in-segment for this LDP MT LSP."
    ::= { mplsLdpMtLspEntry 4 }

mplsLdpMtLspOutSegmentIndex OBJECT-TYPE
    SYNTAX MplsIndexType
    MAX-ACCESS read-only
    STATUS current
    DESCRIPTION
        "Index of out-segment for this LDP MT LSP."
    ::= { mplsLdpMtLspEntry 5 }

mplsLdpMtLspRowStatus OBJECT-TYPE
    SYNTAX Integer32
    MAX-ACCESS read-only
    STATUS current
    DESCRIPTION
        "For creating, modifying, and deleting this row.
        When a row in this table has a row in the active(1)
        state, no objects in this row except this object
        and the mplsLdpMtLspStorageType can be modified."
    ::= { mplsLdpMtLspEntry 6 }

mplsLdpMtLspStorageType OBJECT-TYPE
    SYNTAX StorageType
    MAX-ACCESS read-only
    STATUS current
    DESCRIPTION
        "The storage type for this conceptual row.
        Conceptual rows having the value 'permanent(4)'"
```

need not allow write-access to any columnar
objects in the row."

DEFVAL { nonVolatile }
::= { mplsLdpMtLspEntry 7 }

mplsLdpMtLspOperStatus OBJECT-TYPE

SYNTAX RowStatus

MAX-ACCESS read-only

STATUS current

DESCRIPTION

"The status of this conceptual row. If the value of this
object is 'active(1)', then none of the writable objects
of this entry can be modified, except to set this object
to 'destroy(6)'."

NOTE: if this row is being referenced by any entry in
the mplsLdpLspFecTable, then a request to destroy
this row, will result in an inconsistentValue error."

::= { mplsLdpMtLspEntry 8 }

mplsLdpMtLspService OBJECT-TYPE

SYNTAX QoSService

MAX-ACCESS read-only

STATUS current

DESCRIPTION

"The QoS Service classification index for multiple
topology LSP."

::= { mplsLdpMtLspEntry 9 }

mplsLdpMtConformance OBJECT IDENTIFIER ::= { mplsLdpMtStdMIB 2 }

mplsLdpMtGroups OBJECT IDENTIFIER ::= { mplsLdpMtConformance 1 }

mplsLdpMtEntityGroup OBJECT-GROUP

OBJECTS { mplsLdpMtEntityLastChange, mplsLdpMtEntityIndexNext,
mplsLdpMtEntityMtId, mplsLdpMtEntityAdminStatus,
mplsLdpMtEntityStorageType, mplsLdpMtEntityRowStatus,
mplsLdpMtEntityStatsDiscontinuityTime,
mplsLdpMtEntityStatsSHCOctets, mplsLdpMtEntityStatsDiscards,
mplsLdpMtEntityStatsErrors, mplsLdpMtEntityStatsPackets,
mplsLdpMtEntityStatsOctets }

STATUS current

DESCRIPTION

```

        "Objects that apply to all MPLS LDP MT Entity implementations
."
        ::= { mplsLdpMtGroups 2 }

mplsLdpMtSessionGroup OBJECT-GROUP
    OBJECTS { mplsLdpMtSessionLastChange, mplsLdpMtSessionState,
mplsLdpMtSessionStateLastChange }
    STATUS current
    DESCRIPTION
        "Objects that apply to all MPLS LDP MT Session implementation
s."
        ::= { mplsLdpMtGroups 3 }

mplsLdpMtLspGroup OBJECT-GROUP
    OBJECTS { mplsLdpMtLspLastChange, mplsLdpMtLspIndexNext,
mplsLdpMtLspFecAddr, mplsLdpMtLspFecAddrLength,
mplsLdpMtLspRowStatus, mplsLdpMtLspStorageType,
mplsLdpMtLspOperStatus, mplsLdpMtInSegmentIndex,
mplsLdpMtInSegmentLabelType, mplsLdpMtInSegmentLspType,
mplsLdpMtInSegmentStatsOctets, mplsLdpMtInSegmentStatsPackets,
mplsLdpMtInSegmentStatsErrors, mplsLdpMtInSegmentStatsDiscards,
mplsLdpMtInSegmentStatsSHCOctets,
mplsLdpMtInSegmentStatsDiscontinuityTime,
mplsLdpMtOutSegmentIndex, mplsLdpMtOutSegmentLabelType,
mplsLdpMtOutSegmentLspType, mplsLdpMtOutSegmentStatsOctets,
mplsLdpMtOutSegmentStatsPackets, mplsLdpMtOutSegmentStatsErrors,
mplsLdpMtOutSegmentStatsDiscards,
mplsLdpMtOutSegmentStatsSHCOctets,
mplsLdpMtOutSegmentStatsDiscontinuityTime
    }
    STATUS current
    DESCRIPTION
        "Objects that apply to all MPLS LDP MT LSP implementations."
        ::= { mplsLdpMtGroups 4 }

mplsLdpMtNotificationGroup NOTIFICATION-GROUP
    NOTIFICATIONS { mplsLdpMtLspUp, mplsLdpMtLspDown }
    STATUS current
    DESCRIPTION
        "The notifications for an MPLS LDP MT implementation."
        ::= { mplsLdpMtGroups 5 }

mplsLdpMtCompliances OBJECT IDENTIFIER ::= { mplsLdpMtConformance 2 }

mplsLdpMtModuleFullCompliance MODULE-COMPLIANCE
```

```
STATUS current
DESCRIPTION
    "The Module is implemented with support
    for read-create and read-write.  In other
    words, both monitoring and configuration
    are available when using this MODULE-COMPLIANCE."
MODULE -- this module
    MANDATORY-GROUPS { mplsLdpMtEntityGroup, mplsLdpMtSessionGrou
P,
    mplsLdpMtLspGroup, mplsLdpMtNotificationGroup }
    ::= { mplsLdpMtCompliances 1 }

mplsLdpMtModuleReadOnlyCompliance MODULE-COMPLIANCE
STATUS current
DESCRIPTION
    "The Module is implemented with support
    for read-only.  In other words, only monitoring
    is available by implementing this MODULE-COMPLIANCE"
MODULE -- this module
    MANDATORY-GROUPS { mplsLdpMtEntityGroup, mplsLdpMtSessionGrou
P,
    mplsLdpMtLspGroup, mplsLdpMtNotificationGroup }
    ::= { mplsLdpMtCompliances 2 }

END
```

5. Security Considerations

It needs to be further identified.

6. IANA Considerations

There is no necessary to request new IANA code in the draft.

7. Normative References

- [RFC3813] Srinivasan, C., Viswanathan, A., and T. Nadeau,
"Multiprotocol Label Switching (MPLS) Label Switching
Router (LSR) Management Information Base (MIB)", RFC 3813,
June 2004.

- [RFC3814] Nadeau, T., Srinivasan, C., and A. Viswanathan,
"Multiprotocol Label Switching (MPLS) Forwarding
Equivalence Class To Next Hop Label Forwarding Entry (FEC-
To-NHLFE) Management Information Base (MIB)", RFC 3814,
June 2004.
- [RFC3815] Cucchiara, J., Sjostrand, H., and J. Luciani, "Definitions
of Managed Objects for the Multiprotocol Label Switching
(MPLS), Label Distribution Protocol (LDP)", RFC 3815, June
2004.
- [RFC5036] Andersson, L., Minei, I., and B. Thomas, "LDP
Specification", RFC 5036, October 2007.
- [RFC3410] Case, J., Mundy, R., Partain, D., and B. Stewart,
"Introduction and Applicability Statements for Internet-
Standard Management Framework", RFC 3410, December 2002.
- [I-D.ietf-mpls-ldp-multi-topology]
Zhao, Q., Fang, L., Zhou, C., Li, L., and K. Raza, "LDP
Extensions for Multi Topology Routing", draft-ietf-mpls-
ldp-multi-topology-08 (work in progress), May 2013.

Authors' Addresses

Chen Li
China Mobile
Unit2, Dacheng Plaza, No. 28 Xuanwumenxi Ave, Xuanwu District
Beijing 100053
P.R. China

Email: lichenyj@chinamobile.com

Lianyuan Li
China Mobile
Unit2, Dacheng Plaza, No. 28 Xuanwumenxi Ave, Xuanwu District
Beijing 100053
P.R. China

Email: lilianyuan@chinamobile.com

Lu Huang
China Mobile
Unit2, Dacheng Plaza, No. 28 Xuanwumenxi Ave, Xuanwu District
Xunwu District, Beijing 100053
China

Email: huanglu@chinamobile.com

Tao Chou
Huawei Technology
156 Beiqing Rd
Haidian District, Beijing 100095
China

Email: tao.chou@huawei.com

Quintin Zhao
Huawei Technology
125 Nagog Technology Park
Acton, MA 01719
US

Email: quintin.zhao@huawei.com

Emily Chen
2717 Seville Blvd, Apt 1205
Clearwater, FL 33764
US

Email: emily.chen220@gmail.com

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: January 09, 2014

Z. Li
L. Zheng
Huawei Technologies
July 08, 2013

Mega Label - Expansion of MPLS Label Range
draft-li-mpls-mega-label-00

Abstract

This document describes the requirement scenarios for expansion of MPLS label range. This document also introduce a framework for expansion of MPLS label range-"Mega Label" and the corresponding protocol extensions. This document will update RFC 3032, 5036, 3209 and 3107 if approved.

Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [RFC2119].

Status of This Memo

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

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

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

This Internet-Draft will expire on January 09, 2014.

Copyright Notice

Copyright (c) 2013 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	2
2. Terminology	3
3. Requirement Scenarios	3
3.1. LDP Multi-Topology for MRT FRR	3
3.2. Label Allocation in VPN	3
3.3. Virtual Network Instance	3
4. Framework of Mega Label	4
4.1. Label Stack for Expansion of Label Range	4
4.2. Data Plane	4
4.3. Control Plane	4
5. Protocol Extensions	5
5.1. MPLS LDP	5
5.2. MPLS RSVP-TE	5
5.3. MP-BGP	6
6. IANA Considerations	6
7. Security Considerations	7
8. Acknowledgements	7
9. References	7
9.1. Normative References	7
9.2. Informative References	7
Authors' Addresses	8

1. Introduction

MPLS technology is widely used, but its limited label space which is restricted by the 20bits encoding prevents the MPLS technology from many applications.

This document describe application scenarios that will generate requirements on expansion of MPLS label range. This document also introduce a framework for expansion of MPLS label range - the concept "Mega Label", and the corresponding protocol extensions. This document will update RFC 3032, 5036, 3209 and 3107 if approved.

2. Terminology

LDP MT: LDP Multi-Topology

MRT: Maximally Redundant Trees

3. Requirement Scenarios

3.1. LDP Multi-Topology for MRT FRR

In MRT(Maximally Redundant Trees) FRR scenario([I-D.ietf-rtgwg-mrt-frr-architecture]), when enabling the whole network FRR by incremental deployment of LDP MT in the pure IP network([I-D.li-rtgwg-ldp-mt-mrt-frr]), since the number of internet route is around 500,000, when MPLS labels are allocated in the default topology, blue and red multi-topology simultaneously, the required labels for allocation will reach at least 1.5million. It exceeds the existing MPLS label range dramatically.

3.2. Label Allocation in VPN

In some L3VPN([RFC4364]) deployment, the number of private route already reaches the scale of several ten thousands. The label allocation per Instance method may save the labels to some extent, but it could not be used in some deployment owing to the impossibility of specific flow identification caused by label sharing. Then The label allocation per prefix method maybe have to be used instead and it leads to the required label amount exceeding the existing MPLS label range.

E-VPN([I-D.ietf-l2vpn-evpn]) works in a similar way as L3VPN as to label allocation, but the MAC route could not be aggregated like IP route. This will result in an even worse bottleneck regarding label range for deployment of E-VPN.

3.3. Virtual Network Instance

In NVO3 Scenario([I-D.ietf-nvo3-overlay-problem-statement]), such solutions as VXLAN are introduced to meet the multi-tenant requirement. It extends the number of virtual network instances to 24bits, i.e. 2^{24} . Accordingly, the 2^{20} label range of MPLS is not enough to support the possible virtual network instances.

4. Framework of Mega Label

4.1. Label Stack for Expansion of Label Range

With Mega Label, the label space is extended by stacking MPLS labels as defined in [RFC3032]. As shown in Figure 1, Mega Label consists of Base Label and Remainder Label. The outer layer label for the Mega Label is called as "Base Label", its unit is $M (2^{20})$. The inner layer label for the Mega Label is called "Remainder Label". There could be multiple Base Labels but only one Remainder Label for one Mega Label. If there are multiple Base Labels which represent the label value $N \times M$ and the value of Remainder Label is K , then the value of the Mega Label is $N \times M + K$. M is used as unit of Base Label in this document for acquiring larger label range.

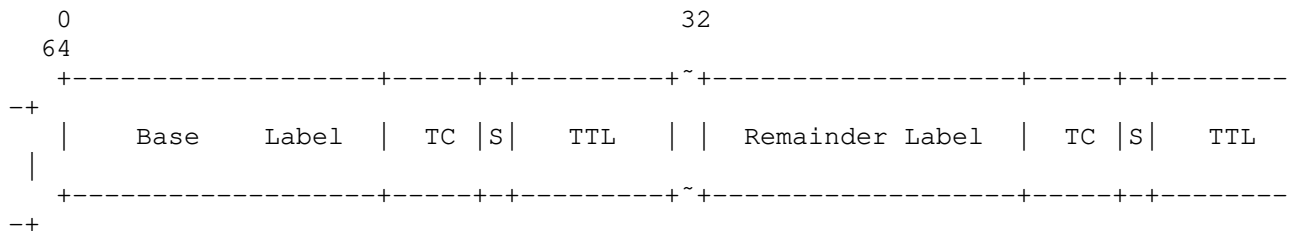


Figure 1: Encapsulation of Mega Label

Base Label could be indicated by the MPLS special labels. For example, special label 5 could be used to indicate the actual base label value $1M$. There are only 16 special labels and the label value 0/1/2/3/7/11/13/14 has already been allocated. Dynamic labels which range is from 16 to $2^{20}-1$ could also be used to indicate a Base Label to solve the above possible limitation proposed by the special label. When a dynamic label is used to indicated a Base Label, it SHOULD be reserved by the downstream LSR for the special usage and no longer be allocated for normal label forwarding.

4.2. Data Plane

When a LSR receives a MPLS packet carrying a Base Label, it SHOULD NOT forward the packet based on this label. Lower layer of label(s) need to be decapsulated and until the Remainder Label is reached. The LSR SHOULD calculate the value of the Mega label based on the actual label value indicated by Base Label(s) and the value of Remainder Label, and then lookup its forwarding table and forward the packet accordingly. The value of EXP/TTL/S field in the Base Label should be consistent with same fields of the lower layer remainder layer. In case of the inconsistency, the value of the EXP/TTL field in the Remainder Label takes precedence.

4.3. Control Plane

MPLS label distribution protocols include LDP, RSVP-TE and MP-BGP etc. These protocols need to be extended to enabling Mega label allocation for one FEC, including Base Label and Remainder Label.

5. Protocol Extensions

5.1. MPLS LDP

The encoding of the LDP Mega Label TLV is as follows, the Mega Label TLV type is to be allocated by IANA. There could be multiple Base Labels and only one Remainder Label in one Mega Label TLV.

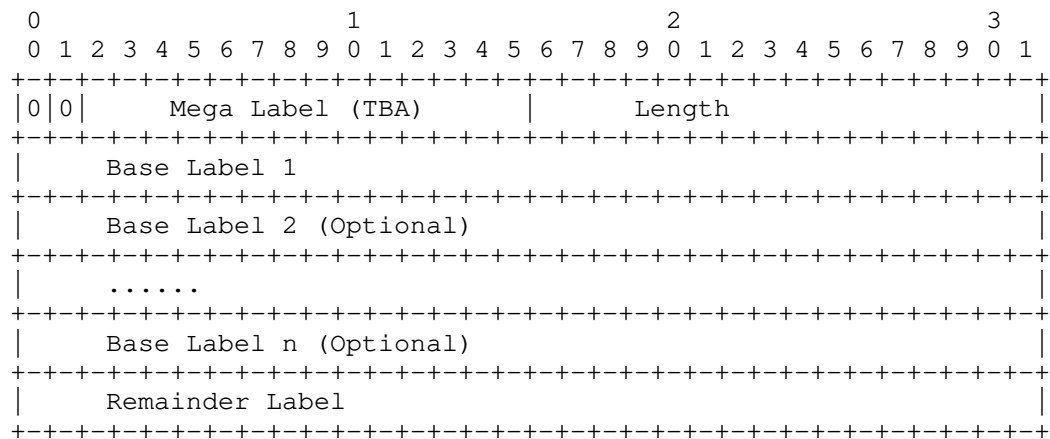
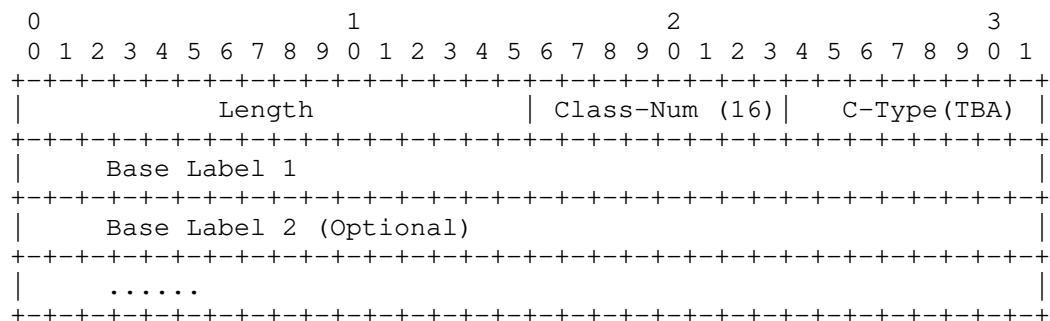


Figure 2. LDP Mega Label TLV

5.2. MPLS RSVP-TE

The encoding of the RSVP-TE Mega Label Object including the common object header is as follows, the C_Type is to be allocated by IANA. There could be multiple Base Labels and only one Remainder Label in one Mega Label Object.



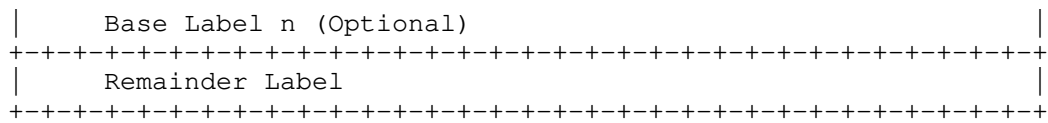


Figure 3. RSVP-TE Mega Label Object

5.3. MP-BGP

The encoding of the MP-BGP Mega Label NLRI is as follows, the label field carries multiple Base Label and only one Remainder Label. Each label is encoded as 3 octets, where the high-order 20 bits contain the label value and the low order bit contains "Bottom of Stack" (as defined in [RFC3032]). The "Bottom of Stack" is cleared for the first and following labels which means it is a Base Label, and the "Bottom of Stack" is set for the last label which means it is a Remainder Label.

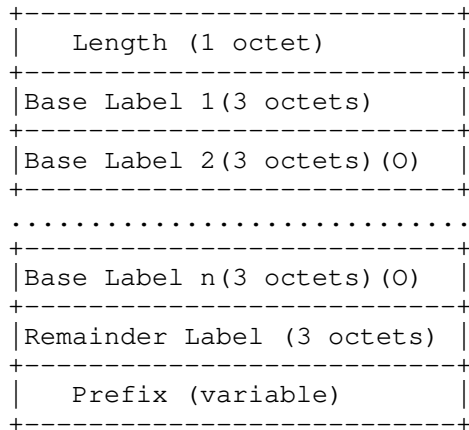


Figure 4. MP-BGP Mega Label NLRI

6. IANA Considerations

The IANA is requested to assign a new TLV from the "TLV Type Name Space " registry.

Value	Meaning	Reference
TBA	Mega Label TLV	this document (sect 4.1)

The IANA is requested to assign a new C-Type from the "Class Type " registry.

Value	Meaning	Reference
-----	-----	-----
TBA	RSVP-TE Mega Label	this document (sect 4.2)

7. Security Considerations

TBD

8. Acknowledgements

The authors would like to thank Loa Andersson for his valuable comments and suggestions on this draft.

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.
- [RFC3031] Rosen, E., Viswanathan, A., and R. Callon, "Multiprotocol Label Switching Architecture", RFC 3031, January 2001.
- [RFC3032] Rosen, E., Tappan, D., Fedorkow, G., Rekhter, Y., Farinacci, D., Li, T., and A. Conta, "MPLS Label Stack Encoding", RFC 3032, 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.
- [RFC5036] Andersson, L., Minei, I., and B. Thomas, "LDP Specification", RFC 5036, October 2007.

9.2. Informative References

- [I-D.ietf-l2vpn-evpn]
Sajassi, A., Aggarwal, R., Henderickx, W., Balus, F., Isaac, A., and J. Uttaro, "BGP MPLS Based Ethernet VPN", draft-ietf-l2vpn-evpn-03 (work in progress), February 2013.
- [I-D.ietf-nvo3-overlay-problem-statement]

Narten, T., Gray, E., Black, D., Fang, L., Kreeger, L., and M. Napierala, "Problem Statement: Overlays for Network Virtualization", draft-ietf-nvo3-overlay-problem-statement-03 (work in progress), May 2013.

[I-D.ietf-rtgwg-mrt-frr-architecture]

Atlas, A., Kebler, R., Envedi, G., Csaszar, A., Tantsura, J., Konstantynowicz, M., White, R., and M. Shand, "An Architecture for IP/LDP Fast-Reroute Using Maximally Redundant Trees", draft-ietf-rtgwg-mrt-frr-architecture-02 (work in progress), February 2013.

[I-D.li-rtgwg-ldp-mt-mrt-frr]

Li, Z., Chou, T., Zhao, Q., and T. Yang, "Applicability of LDP Multi-Topology for Unicast Fast-reroute Using Maximally Redundant Trees", draft-li-rtgwg-ldp-mt-mrt-frr-02 (work in progress), April 2013.

[RFC4364] Rosen, E. and Y. Rekhter, "BGP/MPLS IP Virtual Private Networks (VPNs)", RFC 4364, February 2006.

Authors' Addresses

Zhenbin Li
Huawei Technologies
Huawei Campus, No.156 Beiqing Rd.
Beijing 100095
China

Email: lizhenbin@huawei.com

Lianshu Zheng
Huawei Technologies
Huawei Campus, No.156 Beiqing Rd.
Beijing 100095
China

Email: vero.zheng@huawei.com

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: January 09, 2014

Z. Li
X. Zeng
Huawei Technologies
July 08, 2013

Proxy MPLS Traffic Engineering Label Switched Path(LSP)
draft-li-mpls-proxy-te-lsp-00

Abstract

This document describes a method to setup MPLS TE proxy egress LSP which helps setup end-to-end LSP through stitching MPLS TE proxy egress LSP with BGP LSP in the Seamless MPLS network. The method is achieved by new Proxy Destination Object carried in RSVP-TE messages.

Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [RFC2119].

Status of This Memo

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

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

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

This Internet-Draft will expire on January 09, 2014.

Copyright Notice

Copyright (c) 2013 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	2
2. Terminology	3
3. Solutions	3
4. Proxy Destination Object	4
4.1. Format	4
4.2. Procedures	5
5. IANA Considerations	5
6. Security Considerations	6
7. Acknowledgements	6
8. Normative References	6
Authors' Addresses	6

1. Introduction

Seamless MPLS[I-D.ietf-mpls-seamless-mpls] provides an end to end service independent transport architecture. It removes the need for service specific configurations in network transport nodes. Seamless MPLS uses existing protocols like LDP, IS-IS to build intra-area segments and uses MP-BGP as the inter-area routing and label distribution protocol.

One common procedure of setting up the end-to-end transport LSP is as follows:

1. Setup the access segment LSP from Access Node (AN) to Aggregation Node (AGN) using LDP with longest-match as defined in [RFC5283]. It requires only static routes and it is not necessary to know the actual destination (FEC of the LDP LSP);
2. The Aggregation Node (AGN) stitches the egress LDP LSP with the BGP ingress LSP according to the key of FEC;
3. The remote Aggregation Node (AGN) stitches the egress BGP LSP with an ingress LSP according to the key of FEC.

LSPs set up with MPLS TE (RSVP-TE) provide a higher reliability and better QoS as compared to LSPs set up with LDP. So MPLS TE is always adopted to deploy in the mobile backhaul network. But when the mobile backhaul network integrates with the core/aggregation network based on Seamless MPLS, it is difficult to setup end-to-end MPLS TE

LSP spanning multiple domains. The possible way to setup the end-to-end LSP is that the proxy egress RSVP-TE LSP should be able to setup in the mobile backhaul network to stitch with BGP LSP at the Aggregation Node.

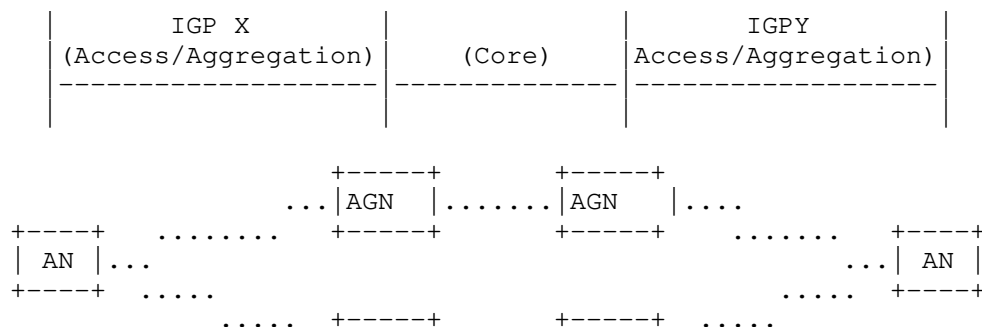
This document extends RSVP-TE extension and introduces a new Proxy Destination Object. Through the RSVP-TE extension the proxy egress LSP can setup for RSVP-TE. It makes possible to setup the end-to-end LSP when deploy MPLS TE in the Seamless MPLS scenario to integrate the mobile backhaul network with the core/aggregation network.

2. Terminology

Proxy Egress LSP: It is defined in Sec. 4.1.4 of [RFC3031]. It is the LSP which is setup by the proxy egress LSR instead of the actual destination LSR.

3. Solutions

When setup a proxy egress RSVP-TE LSP in the Seamless MPLS scenario as shown in the Figure 1, there are two destination addresses to be carried by the RSVP-TE message: the actual destination address is the destination address of the end-to-end LSP by stitching the proxy egress LSP and the BGP LSP, the proxy destination address is the address of Aggregation Node which stitches the proxy egress RSVP-TE LSP and BGP LSP. When setting up the proxy egress RSVP-TE LSP on the Access Node, it is necessary to specify the actual destination address and the proxy destination address. The access node needs to calculate the path based on the proxy destination address for the proxy egress RSVP-TE LSP. The Path message will be sent from the ingress node to the proxy destination node which is identified by the proxy destination address in the message. Then the proxy destination node sends back the Resv message to allocate label and reserve resource. The actual destination address is used to stitch with BGP LSP which has the same address as the actual destination address of the proxy egress RSVP-TE LSP.



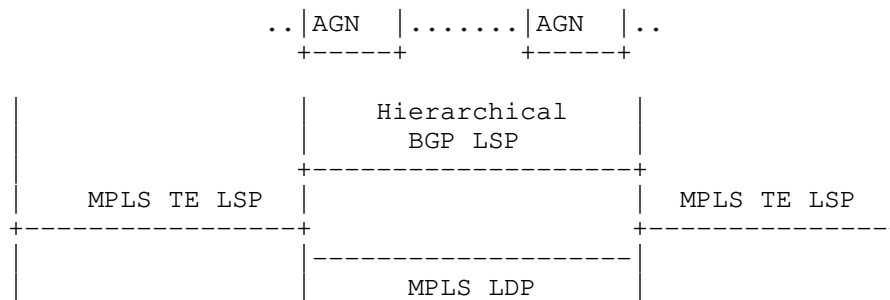


Figure 4 Seamless MPLS Scenario with MPLS TE

In order to support setup of the proxy egress RSVP-TE LSP, the new Proxy Destination Object is introduced to carry the proxy destination address besides that the actual destination address is carried in the Session Object. Both the Session Object and the Proxy Destination Object are carried in the RSVP-TE Path message and Resv message to set up the proxy egress LSP.

4. Proxy Destination Object

4.1. Format

The Proxy Destination Object is an optional object which may be carried in Path or Resv Messages. The Proxy Destination Class-Number is TBD (of form 0bbbbbbb). RSVP-TE routers that do not support the object SHOULD reject the entire message and return an "Unknown Object Class" error.

The format of the Proxy Destination Object is as follows:

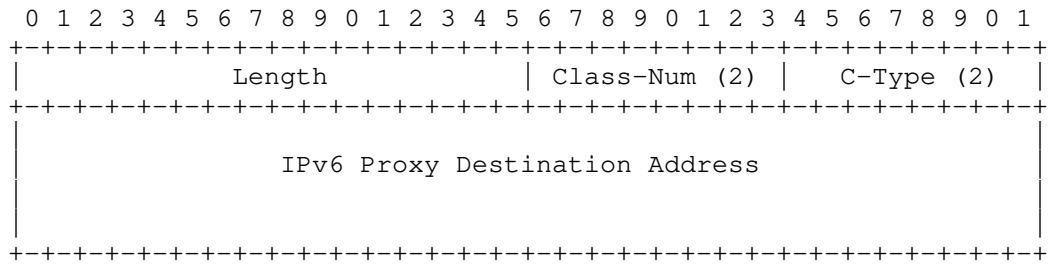
1. IPv4 Proxy Destination Object

```

  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
  +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
  |                               Length                               |
  |                               Class-Num (1)                       |
  |                               C-Type (1)                         |
  |                               IPv4 Proxy Destination Address      |
  +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
  
```

IPv4 Proxy Destination Address: 32 bits. IPv4 address of the proxy destination address of the proxy egress RSVP-TE LSP.

2. IPv6 Proxy Destination Object



IPv6 Proxy Destination Address: 16 bytes. IPv6 address of the proxy destination address of the proxy egress RSVP-TE LSP.

If a message contains multiple Proxy Destination Objects, only the first object is meaningful. Subsequent Proxy Destination Objects SHOULD be ignored and SHOULD NOT be propagated.

4.2. Procedures

When the ingress node sets up the proxy egress LSP, the Proxy Destination Object MUST be inserted in the Path message to indicate the address of the proxy destination node that would stitch the proxy egress LSP with other LSPs. When receive the RESV messages, the ingress node SHOULD check if the Proxy Destination is included. If the Path message include the Proxy Destination object and the corresponding RESV message does not include this object, the ingress node MUST treat the Resv message as wrong messages and MUST NOT set up LSP.

On the transit node, when receiving the messages with Proxy Destination object, it MUST include the Proxy Destination object in the outgoing Path or Resv message without change of the object. When it is necessary for the transit node to calculate the path, the proxy destination address identified by the Object MUST be used instead of the actual destination address identified by the Session Object. If the transit node receives the Path message including the Proxy Destination object but receives the corresponding Resv message which does not include this object, it MUST treat the Resv message as wrong message can MUST NOT set up LSP.

On the egress node, when receiving Path messages with Proxy Destination object, it MUST include this object in the corresponding Resv message.

5. IANA Considerations

TBD.

6. Security Considerations

This document does not introduce any additional security issues above those identified in [RFC3209].

7. Acknowledgements

The authors would like to thank Loa Andersson for his valuable comments and suggestions on this draft.

8. Normative References

- [I-D.ietf-mpls-seamless-mpls]
Leymann, N., Decraene, B., Filsfils, C., Konstantynowicz, M., and D. Steinberg, "Seamless MPLS Architecture", draft-ietf-mpls-seamless-mpls-03 (work in progress), May 2013.
- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC3031] Rosen, E., Viswanathan, A., and R. Callon, "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.
- [RFC5283] Decraene, B., Le Roux, J.L., and I. Minei, "LDP Extension for Inter-Area Label Switched Paths (LSPs)", RFC 5283, July 2008.

Authors' Addresses

Zhenbin Li
Huawei Technologies
Huawei Bld., No.156 Beiqing Rd.
Beijing 100095
China

Email: lizhenbin@huawei.com

Xinzong Zeng
Huawei Technologies
Huawei Bld., No.156 Beiqing Rd.
Beijing 100095
China

Email: zengxinzong@huawei.com

Network Working Group
Internet-Draft
Intended status: Informational
Expires: January 13, 2014

Zhenbin Li
Lei Li
Huawei Technologies
Manuel Julian Lopez Morillo
Vodafone Group Networks
Tianle Yang
China Mobile
July 12, 2013

Seamless MPLS for Mobile Backhaul
draft-li-mpls-seamless-mpls-mbb-00

Abstract

This document introduces the framework of Seamless MPLS to integrate the mobile backhaul network with the core network. New requirements of Seamless MPLS for mobile backhaul networks are defined and corresponding solutions are proposed.

Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [RFC2119].

Status of This Memo

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

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

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

This Internet-Draft will expire on January 13, 2014.

Copyright Notice

Copyright (c) 2013 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	2
2. Terminology	3
3. Scenarios	4
3.1. Scenarios for Network Architecture	4
3.2. Scenarios for Different Edge of Labeled BGP	6
4. Requirements and Solutions	9
4.1. Overall	9
4.2. Scalability	11
4.2.1. Auto Mesh and Enhancement	11
4.2.2. Service-Driven Tunnel	12
4.2.3. Auto Path Computation	12
4.3. Access Stitching	13
4.3.1. Transport Layer Stitching	13
4.3.2. Service Layer Stitching technology	14
4.4. Reliability	16
4.4.1. MRT FRR based on LDP MT	16
4.5. Policy Control	17
4.6. OAM	17
4.6.1. L3VPN PM	17
4.6.2. IPFPM (IP Flow Performance Measurement)	18
4.6.3. Service Path Visualization	18
5. IANA Considerations	18
6. Security Considerations	18
7. Acknowledgements	19
8. Normative References	19
Authors' Addresses	20

1. Introduction

Seamless MPLS [I-D.ietf-mpls-seamless-mpls] describes an architecture which can be used to extend MPLS networks to integrate access and

core/aggregation networks into a single MPLS domain. It provides a highly flexible and a scalable architecture and the possibility to integrate 100.000 of nodes. One of the key elements of Seamless MPLS is the separation of the service and transport plane: it can reduce the service specific configurations in network transport node.

The main purpose of Seamless MPLS is to deal with the integration of access networks and core/aggregation networks. The typical access devices taken into account are DSLAM(Digital Subscriber Link Access Multiplexer), etc. Now the mobile backhaul service has been deployed widely, the requirement of the integration of mobile backhaul networks and core networks has been proposed. Though some approaches of Seamless MPLS can be reused for the integration, there has to be some new issues to be dealt with when integrate these networks based on MPLS technologies.

This document describes new requirements and solutions for Seamless MPLS to extend the core domain to integrate the mobile backhaul networks. It can enable a flexible deployment of an end to end mobile backhaul service delivery. Though the Seamless MPLS approach for the integration of the core network and the mobile network tries to use existing and well known protocols, some new features or protocol extensions have to be introduced for the integrated network to provide the unified service.

Currently, this document focuses on end to end unicast LSP. Multicast will be described in the future version.

2. Terminology

This document uses the following terminology:

- o ABR: Area Border Router
- o ASBR: AS Border Router
- o ASG: Aggregation Site Gateway
- o CSG: Cell Site Gateway
- o LFA: Loop Free Alternate
- o NPE: Network Provider Edge
- o PE: Provider Edge
- o RNC: Radio Network Controller

- o RSG: RNC Site Gateway
- o SPE: Switching Provider Edge
- o UPE: Under Provider Edge

3. Scenarios

Existing mobile backhaul networks have different topology and network architectures composed by devices with variable capability . Seamless MPLS should be able to adapt to different scenarios to support the integration of mobile backhaul networks.

3.1. Scenarios for Network Architecture

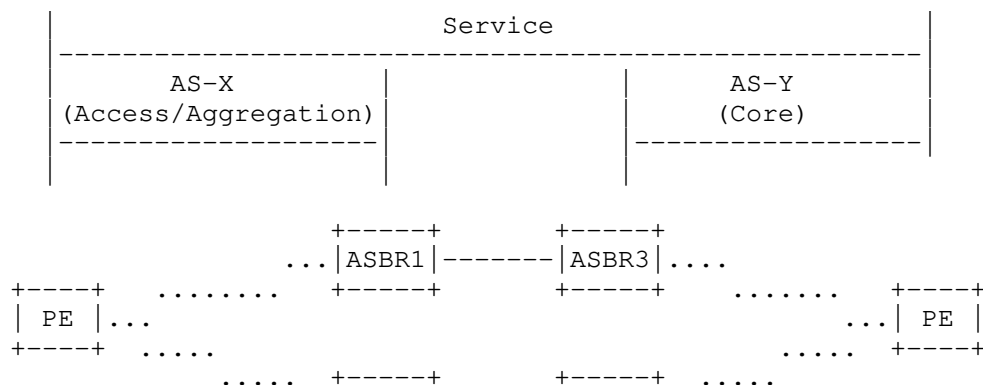
Mobile backhaul networks are usually built using hierarchical network structure with access network, aggregation network and core network. These networks are always separated by AS or IGP area. Along with the progress of network integration, the integration network can be summarized as following scenarios.

Network Architecture 1: Network separated by ASes

In current networks it is common that the core network and the mobile backhaul network have different AS numbers. The core network usually uses a public AS number for internet connection. On the other hand the mobile backhaul network including access network and aggregation network usually uses a private AS number just for local services. So the integrated end-to-end service means to across different ASes.

Scenario 1: ASes connected by different ASBRs

This is the most common scenario. In this scenario there are redundant ASBRs in each AS to connect with the other AS back to back.



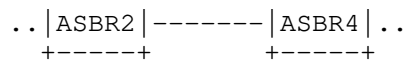


Figure 1 Redundant ASBRs connected Back to Back

The transport layer of Seamless MPLS for this scenario is the same as the Option C Inter-AS VPN scenario defined by [RFC4364]. In this scenario, Seamless MPLS uses label distribution enabled IBGP and EBGp to establish the end-to-end BGP LSP to support services (using IPv4 as the example).

- o IBGP distributes the PE's 32/ route to ASBRs in source AS (P devices need not know the PE's 32/ route).
- o EBGp redistributes labeled IPv4 routes from source AS to neighboring AS.
- o IBGP distributes the PE's 32/ route from ASBRs to ingress PEs in target AS.

Scenario 2: ASes connected by integrated ASBRs

In this scenario there are still redundant ASBRs in each AS. But these ASBRs integrates together to reduce a pair of devices. This scenario can effectively reduce the number of devices and costs. Other devices in each AS such as PEs and Ps need not be impacted.

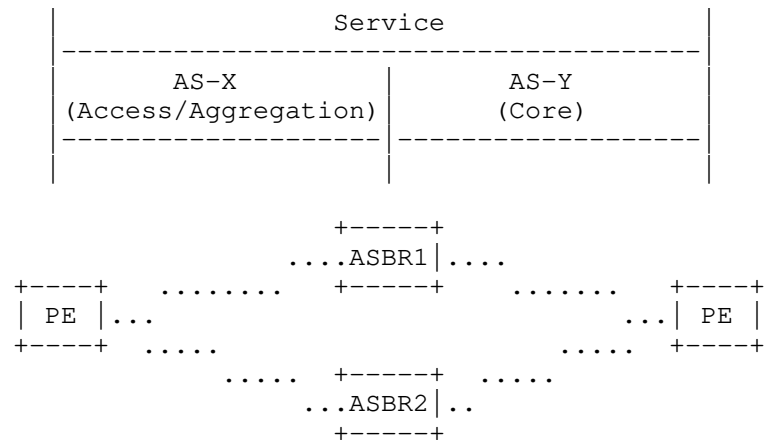


Figure 2 Integrated ASBRs

The transport layer of Seamless MPLS for this scenario is similar with the network architecture 1 (using IPv4 as the example).

- o The same in each AS domain. Still use IBGP to distribute the PE's 32/ routes.
- o The integrated ASBR should support two different ASes and can redistribute the labeled IPv4 routes from one AS to neighboring AS.

Network Architecture 2: Different network integrated in one AS but separated by different IGP areas

This scenario is far different from most of current mobile backhaul networks. Core network is deployed in a same AS with access/aggregation network. The core network, aggregation network and access network are just separated by IGP areas for scalability.

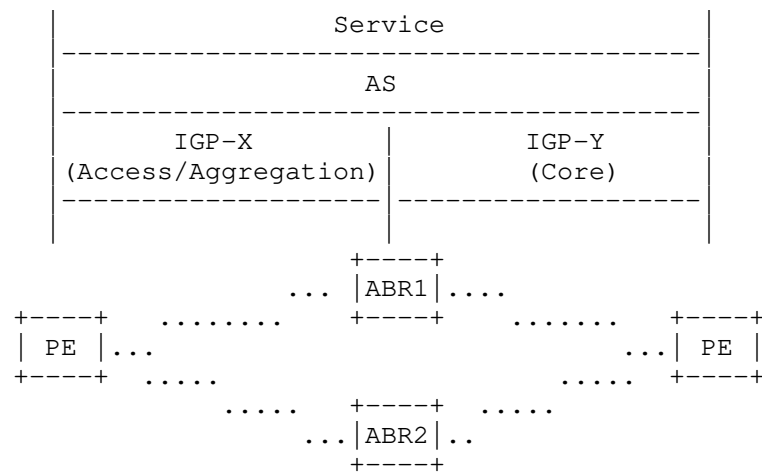


Figure 3 Integrated network in one AS

In this scenario, Seamless MPLS uses labeled IBGP to establish the end-to-end BGP LSP to support services.

3.2. Scenarios for Different Edge of Labeled BGP

Devices in existing mobile backhaul networks vary in capacity. Labeled BGP capability may not be able to be supported by all devices, especially by lower level nodes in access network. Seamless MPLS based on labeled BGP technology should adapt to different situations. Based on the location of the edge of labeled BGP, there should be three possible scenarios.

Scenario 1: Cell Site/User PE devices as the edge of labeled BGP

The transport layer in this scenario should be totally end-to-end BGP LSP. The scenario requires the ingress PE(access devices) to encapsulate a three-label stack on the packet. This requirement maybe difficult to be satisfied by all kinds of access devices, especially access devices with very low capacity.

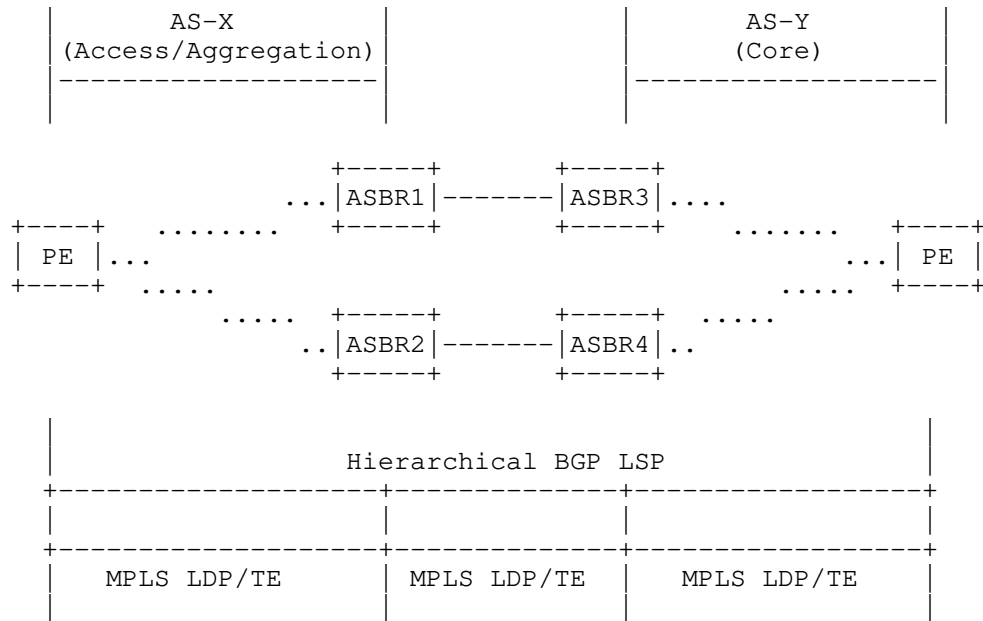
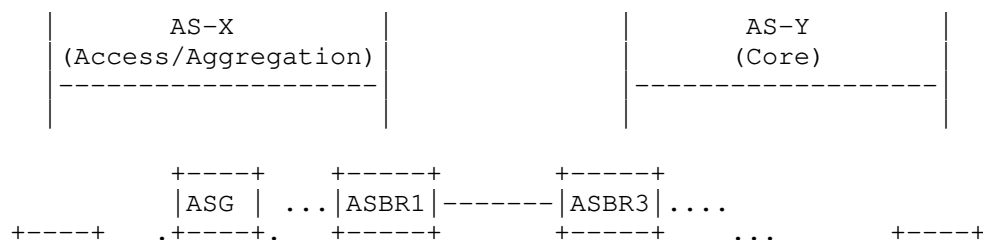


Figure 4 Labeled BGP ended at access devices

Scenario 2: ASG nodes as the edge of labeled BGP

In this scenario, access nodes (PEs) directly connected with eNodeB can not support labeled BGP. Access nodes only support basic MPLS functionality with basic route functionality using static or default routes. ASG devices should stitch MPLS LDP/TE LSP in access network and BGP LSP in aggregation/core network to support end-to-end services.



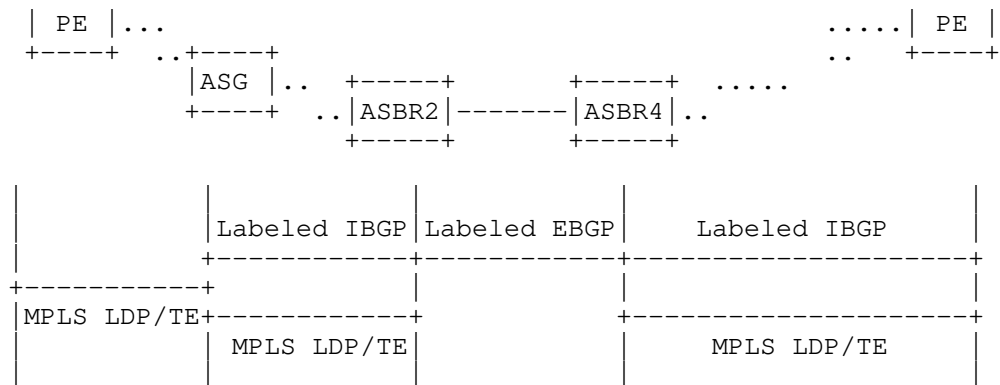


Figure 5 Labeled BGP ended at ASG(P) devices

Scenario 3: RSG(ASBR) devices as the edge of labeled BGP

In this scenario devices in the access and aggregation network just support basic MPLS functionality. ASBR nodes should stitch MPLS LDP/TE LSP in access/aggregation network and BGP LSP in core network for end-to-end service across different domains.

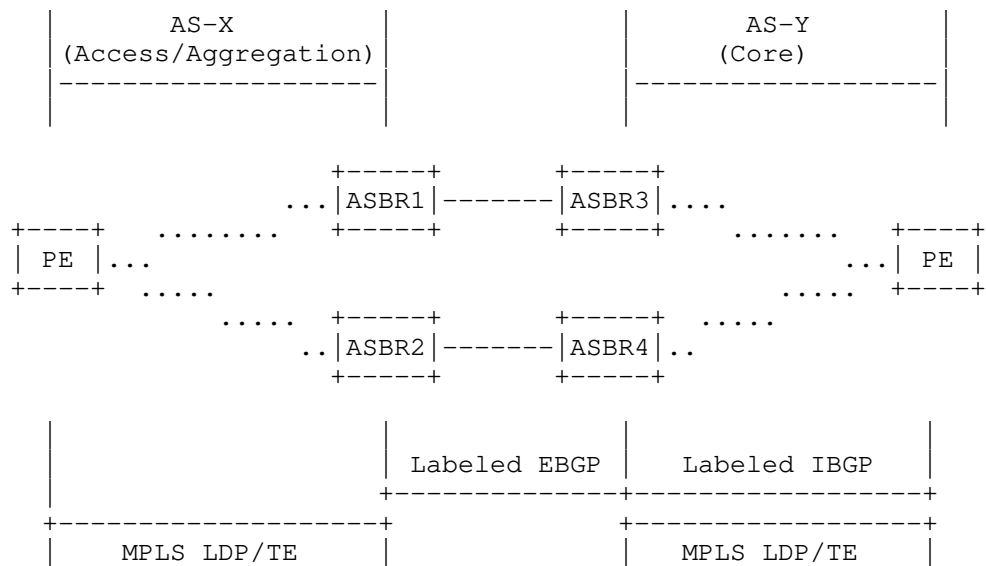


Figure 6 Labeled BGP ended at ASBRs

4. Requirements and Solutions

4.1. Overall

The typical mobile backhaul network is shown in the figure 1. It usually adopts ring topology to save fiber resource and it is divided into the aggregate network and the access network. Cell Site Gateways (CSGs) connects the eNodeBs and RNC Site Gateways (RSGs) connects the RNCs. The mobile traffic is transported from CSGs to RSGs.

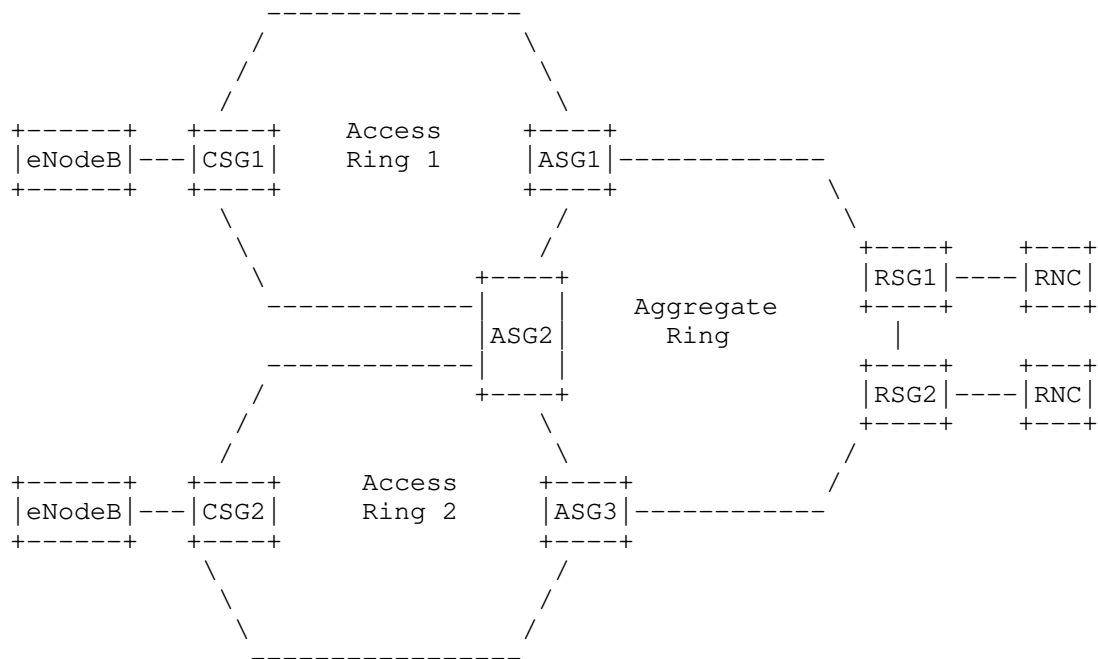


Figure 7 Mobile Backhaul Network

Seamless MPLS [I-D.ietf-mpls-seamless-mpls] defines the possible requirements for the integration of the access network and the aggregation/core network. In the mobile backhaul network, being different from the typical access devices (DSLAM, MSAN), CSGs and RSGs of the mobile backhaul network needs to support rich MPLS features such as path design, protection switch, OAM, etc., to provide SDH-like service. On the other hand, CSGs have the same scalability limitations as access devices. Seamless MPLS for mobile backhaul has to take into account the difference and the similarity and new requirements are proposed.

1. Requirements on MPLS TE

In the mobile backhaul network, in order to provide SDH-like service, MPLS TE or MPLS TP technologies are always introduced besides MPLS LDP. When integrate the core network and the mobile backhaul network, the interworking between IGP/BGP and RSVP-TE is inevitable. There are following requirements:

-- Proxy Egress LSP and BGP LSP Stitching: When the end-to-end LSP setup in the Seamless MPLS domain, MPLS TE LSP in the mobile backhaul should be stitched with the BGP LSP in the core network. In addition, since the end-to-end MPLS TE LSP cannot setup spanning multiple areas, the proxy egress RSVP-TE LSP should be able to setup in the mobile backhaul network.

-- OAM: There are complete OAM solutions for MPLS TE/MPLS-TP including fault management (FM) and performance monitoring (PM). For IGP/BGP, the OAM mechanism, especially the performance monitoring mechanism is not enough. Thus the end-to-end OAM for the mobile backhaul service can not be provided. The Seamless MPLS architecture should propose unified OAM mechanisms to satisfy the requirements of the end-to-end services.

-- Protection: The protection switch mechanism has been provided in the mobile backhaul network to achieve convergence in 50ms. When integrate the core network, the end-to-end convergence in 50 ms should be guaranteed.

-- Scalability: Comparing with MPLS LDP, there exists more scalability issues for MPLS TE. Though the network scale can be controlled by dividing the network into multiple IGP areas in the Seamless MPLS domain, there is more complex configuration for MPLS TE than MPLS LDP since the rich MPLS TE properties should be specified. The provision issue has much negative effect on the scalability of MPLS-TE-based network.

2. Requirements on Ring Network

In mobile backhaul network, CSGs always access the ring network. The approaches proposed in the [I-D.ietf-mpls-seamless-mpls] face challenges in the ring network.

-- LDP DoD: There exists multiple nodes from a CSG to an ASG in the ring network. This means label request messages and label mapping messages should be sent through multiple nodes for LDP DoD. If static routes are used, there are a great deal of routes which should be configured statically in the network. This provision is hard to be accepted by the service provider.

-- LFA(Loop Free Alternate): The route loop will be bound to happen in the ring network. This means that the backup route does not exist in specific nodes of the ring network according to LFA. If LDP is used and convergence in 50 ms must be guaranteed for the mobile backhaul service, the new FRR solution must be proposed to cover the whole network.

3. Requirements on L3VPN

FMC(Fixed Mobile Convergence) is being taken into account for the mobile backhaul network. In order to achieve higher scalability, L3VPN is provisioned to bear the mobile backhaul service besides PW. There are following requirements:

-- Policy control: The routes in the L3VPN can be advertised in a larger scope comparing with the point-to-point PW setup in the L2VPN. Considering the limited capability of the access devices (CSGs), complex policy control on route advertisement has to be introduced. This cause the provision becomes complex and error-prone. Simplifying the policy control is mandatory in the Seamless MPLS domain.

-- L3VPN OAM: There exists mature OAM mechanism based on PW for L2VPN. The similar mechanism should be provided for the L3VPN to satisfy the SDH-like OAM requirement for the mobile backhaul service.

4.2. Scalability

[I-D.li-mpls-serv-driven-co-lsp-fmwk] describes the massive configuration issue for MPLS TE. As the mobile backhaul service develops and the network scale expands, more and more LTE eNodeBs and associated Cell Site Gateways(CSGs) are added in the networks to connect the RNCs and associated RNC site gateways(RSGs). This proposes the requirement of a great deal of MPLS TE tunnels which connect CSGs and RSGs. Calculated using the typical MPLS TE tunnel configuration, there will be hundreds of thousands of command lines need to be configured for MPLS TE. In addition, the return path issue for BFD for LSP will deteriorate the configuration work since the configuration is necessary to guarantee the return the path is the same as the forward path for MPLS TE tunnels. Comparing with LDP, the provision work for MPLS TE is time-consuming and error-prone which has much negative effect on the scalability. When Seamless MPLS is applied to the mobile backhaul network, the scalability of MPLS TE must be improved by effective solutions.

4.2.1. Auto Mesh and Enhancement

[RFC4972] proposes an automatic discovery mechanism to discover the set of LSR members of a mesh in order to automate the creation of mesh of TE LSPs. Through Interior Gateway Protocol (IGP) routing extensions, the LSRs members of a specific mesh can be discovered. Then the LSR can trigger setup of MPLS TE tunnels to other LSRs of the same mesh. [I-D.li-ccamp-role-based-automesh] proposes the optimization for the auto mesh mechanism. In some application scenarios, it is not necessary to setup full mesh MPLS TE tunnels. The optimized auto mesh mechanism is to not only advertise the membership of a mesh, but also advertise the role of the member. Thus the MPLS TE tunnel can setup based on both the membership and the role. It can save the unnecessary setup of MPLS TE tunnels.

4.2.2. Service-Driven Tunnel

[I-D.li-mpls-serv-driven-co-lsp-fmwk] proposes the service-driven mechanism and framework for setup of MPLS TE tunnels. LDP LSP has advantage over MPLS TE in scalability since LDP LSP setup is topology-driven which is a scalable way to adapt to the large-scale network. On the other hand, MPLS TE LSP is always setup to bear specific services such as L3VPN and L2VPN. That is, MPLS TE LSPs will not be setup aimlessly which is always inevitable for MPLS topology-driven LSP if there is no policy exerted on it. So the service-driven mechanism is proposed to trigger MPLS TE LSP setup automatically by the service instead of explicitly configuring each tunnel and its traffic engineering attributes. The service-driven method also has much advantage in the process of setting up co-routed TE LSPs. The mobile backhaul service transported by MPLS TE LSPs is always bi-directional. The characteristic can be utilized to setup the forward MPLS TE LSP and the co-routed reverse MPLS TE LSP. The details of the procedures can refer to [I-D.li-mpls-serv-driven-co-lsp-fmwk].

4.2.3. Auto Path Computation

No matter the auto mesh mechanism or the service-driven mechanism is used to automate the creation of MPLS TE LSPs, several set of MPLS TE properties need to be specified for these MPLS TE tunnels. If a group of tunnels can share one set of MPLS TE properties, they can be triggered to setup automatically. On the contrary, if there are distinct MPLS TE properties for MPLS TE tunnels, the configuration work can not be saved since even if the auto tunnel mechanism is adopted the MPLS TE properties has to be configured for each tunnel which is like the current explicit configuring of traffic engineering properties of a tunnel. For MPLS TE, the explicit path is such a tough thing which can have negative effect on the auto tunnel mechanism. [I-D.li-ospf-auto-mbb-te-path] describes the scenarios of the mobile backhaul service in which the explicit path has to be

used. Accordingly TA (TE Area) and TL(TE Layer) are introduced for the design of MPLS TE network view. Thus less MPLS TE properties need to be specified for MPLS TE tunnels and automation of path computation can be improved. As a result the auto tunnel mechanism can be applied to improve the scalability of MPLS TE.

4.3. Access Stitching

To reduce the requirement on lower level network devices(access nodes/ ASG nodes, etc.) and keep these devices as simple as possible, the MPLS stitching technology should be deployed at the edge of labeled BGP nodes. Thus nodes under the stitching points just need to support basic MPLS function with IGP or even just static routes. The position of the stitching point has been discussed in section 3.2. This section introduces the stitching solutions.

1. Transport layer stitching. This kind of stitching technology ensures the transport LSP can setup end to end. The service is just to deploy at the end points and the transit nodes in network need not perceive services.
2. Service layer stitching. This kind of stitching technology establishes a hierarchical service end to end. This technology can reduce the load of service session on access nodes. But the stitching nodes need to perceive services.

4.3.1. Transport Layer Stitching

Transport layer stitching technology can stitch multi-segment transport LSPs together to establish an end-to-end LSP.

4.3.1.1. Proxy TE

In mobile backhaul network MPLS TE has been adopted to provide SDH-like service. When the end-to-end LSP setup in the Seamless MPLS domain, MPLS TE LSP in the mobile backhaul network should be stitched with the BGP LSP in the core network. [I-D.li-mpls-proxy-te-lsp] proposes the solution to setup proxy egress LSP by RSVP-TE. When setup such LSP, there are two addresses will be carried by RSVP-TE Path/Resv messages: the actual destination address and the proxy node address. RSVP-TE Path message will be sent along the path to the proxy node instead of the actual destination node. When Path messages arrives at the proxy node, it will send back Resv message to allocate label and reserve resource. The proxy node can use the actual destination address advertised by the Path message to stitch with BGP LSP. With Proxy TE, the path for the LSP is calculated with the proxy node as the destination instead of its actual destination. So the MPLS TE information related with the path to the actual

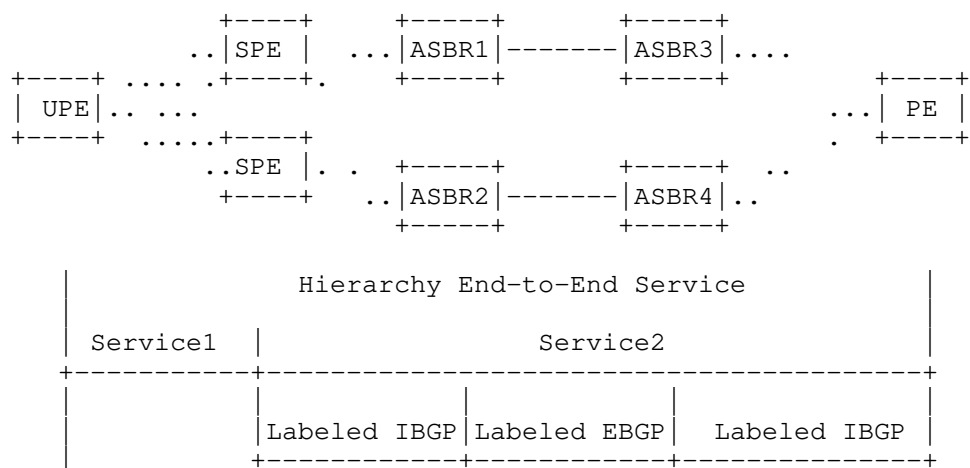
destination is not necessary to flood in the mobile backhaul network. This reduces the state maintenance and process burden of access nodes.

4.3.1.2. Proxy LDP DoD

If LDP DoD is adopted for Seamless MPLS in the mobile haul network, since there are multiple hops from the CSG to ASG or RSG, it is troublesome to configure static routes to a specific destination on all nodes. Like Proxy TE, LDP can also setup proxy egress LSP by specifying the proxy node. When setup such LSP, there are two addresses will be carried by LDP Label Request/Label Mapping messages: the actual destination address and the proxy node address. LDP Label Request message will be sent along the path to the proxy node instead of the actual destination node. When Label Request messages arrives at the proxy node, it will sent back Label Mapping message to allocate label. The proxy node can use the actual destination address advertised by the Label Request message to stitch with BGP LSP. With Proxy LDP, the route for the proxy node will be used to setup LSP for the actual destination. So the static routes for the actual destination are not necessary in the mobile backhaul network. This reduces the route number and process burden of access nodes.

4.3.2. Service Layer Stitching technology

There is another stitching technology, which is not only just stitching transport layer but also stitching the service layer to help establish services across different domains. Service layer stitching technology should coexist with common services in an MPLS network.



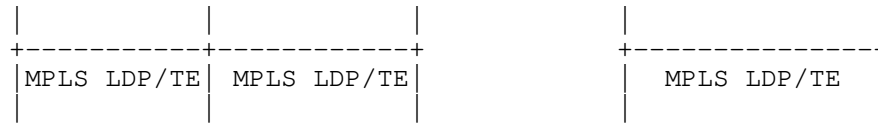


Figure 8 Architecture of Service Layer Stitching

The general architecture of service layer stitching is shown in figure 8. The service stitching technologies are related with different service types.

4.3.2.1. L3 Service Stitching - Hierarchy of VPN (HoVPN)

Hierarchy of VPN (HoVPN) can stitching two segment VPN in one domain together to establish a VPN service across different domains and simplify the process of access devices (AN/CSG).

1. Process of Control Plane

- o UPEs provide the access service for users. These UPEs act as CSG/AN to maintain the VPNv4 routes of the directly connected VPN sites and just default VPNv4 routes advertised from SPEs. For transport layer UPEs just maintain LSPs to SPEs instead of establishing Hierarchical LSPs to all remote PEs. For service layer UPEs just establish VPNv4 MP-BGP session with SPEs instead of establishing them with all remote PEs.
- o SPEs maintain all VPNv4 routes of the VPN sites connected through the UPEs and PEs, including the routes from the local and the remote sites. Instead of advertising routes from the remote sites to the UPEs, the SPE only advertises the default route carrying label to the UPEs for the specific VPN instance. SPEs establish hierarchy LSPs through labeled BGP to all other PEs and basic LSPs /CR-LSPs through MPLS LDP/RSVP-TE to UPEs.

2. Process of Forwarding Plane

- o UPEs stack two layer labels for packets to SPEs. The bottom label is assigned by the SPE and it is associated with the default route in a particular VRF. The top label is assigned by the UPE's IGP Next Hop, which is associated with to the /32 route to the SPE.
- o When packets from UPEs arrive at SPEs, the bottom label associated with the default route in a particular VRF will introduce the packet to loop up the forwarding entry in corresponding VRF. SPEs would send packets to remote PEs with three layer label stack. SPEs will swap the bottom label to a new bottom label which is

assigned by remote PE. The middle label is assigned by the ASBR, which is associated with the /32 route to the remote PE. The top label is assigned by the SPE's IGP Next Hop, corresponding to the /32 route to the ASBR.

In HoVPN, since the SPE will only advertise the label binding with the default route for a specific VRF to the UPE, it can greatly reduce the route load and the process burden in the UPE.

4.3.2.2. L2VPN Service stitching- Multi-Segment PW

Multi-Segment PW(MS-PW) can stitching two segment PW in one domain together to establish an end-to-en PW across different domains. With MS-PW, it is not necessary to setup LDP remote sessions among the UPEs and the remote PEs. It is just to setup LDP remote session among UPEs and SPEs. Thus it can reduce the load and the process burden proposed by LDP remote sessions.

1. Process of Control Plane

- o UPEs provide the access service for users. These UPEs act as CSG/AN to establish LDP remote sessions with SPEs instead of establishing sessions with all remote PEs.
- o SPEs establish SS-PW(Single-Segment PW) with UPEs and remote PEs and stitching(switch) these two SS-PW together to establish hierarchy PW services across different domains.

2. Process of Forwarding Plane

The detail of process of the forwarding plane refers to [I-D.ietf-pwe3-ms-pw-requirements].

4.4. Reliability

4.4.1. MRT FRR based on LDP MT

[I-D.li-rtgwg-ldp-mt-mrt-frr] describes the scenarios of LDP Multi-topology(MT) for unicast fast-reroute using Maximally Redundant Trees(MRT). MRT FRR can provide 100% coverage for fast-reroute of unicast traffic and LDP multi-topology has been proposed to provide multi-topology-based unicast forwarding in the architecture. Combining MRT FRR and LDP MT can be easy to achieve the object of high reliability in IGP/LDP networks. Ring topology has been widely adopted in mobile backhaul networks. The route loop will be bound to happen in the ring network and the existing LFA technology cannot cover the whole network to implement fast reroute functionality. MRT FRR based on LDP MT can be adopted in the mobile backhaul network

using LDP to provide 100% coverage for fast-reroute. And the solution can also be deployed uniformly in the core network using LDP to achieve high scalability.

4.5. Policy Control

BGP as a route protocol for inter-AS now is used for Seamless MPLS to establish end-to-end hierarchical LSP or deploy VPN services. BGP route policy based on IP-Prefix or communities are usually used to control the path. The design and configuration is complex and error-prone. In fact, BGP in Seamless MPLS is used to propagate labeled BGP routes across different domains to implement network convergence. This means several contiguous BGP networks are under the uniform administration. It is not like the traditional BGP networks which may be under the administration of different service providers. In this cases, thinking on the security of the network can be reduced. On the contrary, when advertise routes in the Seamless MPLS domain, it is desirable for BGP to carry more information to help select routing more intelligently. It can reduce the cost proposed by complex policy control design and be able to adapt to network change easily.

4.6. OAM

Mobile Backhaul is a sensitive network on latency timer, packet loss rate, performance and so on. Therefore, unified OAM mechanism is necessary to ensure the end-to-end network management including fault and performance management.

OAM mechanisms is complete and mature for L2VPN and MAC services. However, L3VPN are introduced in the mobile backhaul network for better scalability. The OAM mechanisms for IP and L3VPN is not sufficient to satisfy the OAM requirement of the mobile service, especially for performance monitoring. On the other hand, the Seamless MPLS is in fact composed by multiple contiguous networks. More convenient mechanisms should be introduced for maintenance of the end-to-end path.

4.6.1. L3VPN PM

FMC becomes the requirement of mobile backhaul network and L3VPN is introduced to get better scalability for service provisioning. Unlike existing mature OAM mechanism based on PW for L2VPN, L3VPN lacks of similar mechanisms to satisfy the SDH-like OAM requirement for the mobile backhaul service. [I-D.zheng-l3vpn-pm-analysis] analyzes the difficulty to implement performance monitoring for L3VPN owing to its MP2P or MP2MP service models. [I-D.dong-l3vpn-pm-framework] proposes the framework to implement

L3VPN PM. The point-to-point connection between two VRFs, called as VRF-to-VRF tunnel, is established. Thus the existing MPLS OAM mechanisms based on P2P LSP models can be reused for L3VPN.

4.6.2. IPFPM (IP Flow Performance Measurement)

It is lack of effective performance monitoring mechanisms for IP-based mobile service. The existing PM mechanism can not guarantee that the path of the injected OAM packets is same with the real traffic. If out-of-order arrival of packets happens, the accuracy of the measurement result will be affected negatively for the IP service.

[I-D.chen-coloring-based-ipfpm-framework] defines a new performance measurement mechanism for Service Level Agreement (SLA) verification and trouble shooting (e.g., fault localization or fault delimitation) named as IP Flow Performance Measurement (IPFPM). This measurement mechanism can measure performance based on 5-tuple encapsulation information of IP flow without injecting any extra OAM packet to the flow. The accuracy of the measurement result can be guaranteed even if out-of-order arrival of packets happens by setting one unused bit of the IP header of packets to "color" the packets into different color blocks. The solution can adapt to various IP based network architecture (pure IP, L3VPN, etc.)

4.6.3. Service Path Visualization

Seamless MPLS provides an architecture to support end-to-end services across multi-separated IP/MPLS domains. Existing path detect technologies (e.g. IP/LSP Ping and Trace) can only trace the path in different layers or different network segments. So it is ineffective to get the end-to-end path combining these technologies for maintenance of the network. On the other hand, existing technologies do not encapsulate the same 5-tuple {source IP address, destination IP address, source port number, destination port number, IP protocol number} as the real traffic. This means the path maybe be different between the OAM packets and the real flow's packets when there are more than one outgoing paths and the forwarding decision is determined by hash based on 5-tuple information in the IP packet. According to these new requirements, the new solution should be introduced to maintain the end-to-end path more conveniently.

5. IANA Considerations

This document makes no request of IANA.

6. Security Considerations

TBD.

7. Acknowledgements

The authors would like to thank Loa Andersson for his valuable comments and suggestions on this draft. The authors would also like to acknowledge the important contributions of Yuanbin Yin on IPFPM and Service Path Visualization.

8. Normative References

[I-D.chen-coloring-based-ipfpm-framework]

Chen, M., Liu, H., and Y. Yin, "Coloring based IP Flow Performance Measurement Framework", draft-chen-coloring-based-ipfpm-framework-01 (work in progress), February 2013.

[I-D.dong-l3vpn-pm-framework]

Dong, J., Li, Z., and B. Parise, "A Framework for L3VPN Performance Monitoring", draft-dong-l3vpn-pm-framework-01 (work in progress), April 2013.

[I-D.ietf-mpls-seamless-mpls]

Leymann, N., Decraene, B., Filsfils, C., Konstantynowicz, M., and D. Steinberg, "Seamless MPLS Architecture", draft-ietf-mpls-seamless-mpls-03 (work in progress), May 2013.

[I-D.ietf-pwe3-ms-pw-requirements]

Bocci, M. and L. Martini, "Requirements for Multi-Segment Pseudowire Emulation Edge-to-Edge (PWE3)", draft-ietf-pwe3-ms-pw-requirements-07 (work in progress), June 2008.

[I-D.li-ccamp-role-based-automesh]

Li, Z. and M. Chen, "Routing Extensions for Discovery of Role-based MPLS Label Switching Router (MPLS LSR) Traffic Engineering (TE) Mesh Membership", draft-li-ccamp-role-based-automesh-00 (work in progress), February 2013.

[I-D.li-mpls-proxy-te-lsp]

Li, Z. and X. Zeng, "Proxy MPLS Traffic Engineering Label Switched Path(LSP)", draft-li-mpls-proxy-te-lsp-00 (work in progress), July 2013.

[I-D.li-mpls-serv-driven-co-lsp-fmwk]

Li, Z. and J. Dong, "A Framework for Service-Driven Co-Routed MPLS Traffic Engineering LSPs", draft-li-mpls-serv-driven-co-lsp-fmwk-01 (work in progress), April 2013.

[I-D.li-ospf-auto-mbb-te-path]

Li, Z., Zhang, L., and Y. Liu, "OSPF Extensions for Automatic Computation of MPLS Traffic Engineering Path Using Traffic Engineering Layers and Areas", draft-li-ospf-auto-mbb-te-path-00 (work in progress), February 2013.

[I-D.li-rtgwg-ldp-mt-mrt-frr]

Li, Z., Chou, T., Zhao, Q., and T. Yang, "Applicability of LDP Multi-Topology for Unicast Fast-reroute Using Maximally Redundant Trees", draft-li-rtgwg-ldp-mt-mrt-frr-02 (work in progress), April 2013.

[I-D.zheng-l3vpn-pm-analysis]

Zheng, L., Li, Z., and B. Parise, "Performance Monitoring Analysis for L3VPN", draft-zheng-l3vpn-pm-analysis-01 (work in progress), April 2013.

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

[RFC4364] Rosen, E. and Y. Rekhter, "BGP/MPLS IP Virtual Private Networks (VPNs)", RFC 4364, February 2006.

[RFC4972] Vasseur, JP., Leroux, JL., Yasukawa, S., Previdi, S., Psenak, P., and P. Mabbey, "Routing Extensions for Discovery of Multiprotocol (MPLS) Label Switch Router (LSR) Traffic Engineering (TE) Mesh Membership", RFC 4972, July 2007.

Authors' Addresses

Zhenbin Li
Huawei Technologies
Huawei Bld., No.156 Beiqing Rd.
Beijing 100095
China

Email: lizhenbin@huawei.com

Lei Li
Huawei Technologies
Huawei Bld., No.156 Beiqing Rd.
Beijing 100095
China

Email: lily.lilei@huawei.com

Manuel Julian Lopez Morillo
Vodafone Group Networks
Parque Empresarial Castellana Norte. Isabel Colbrand 22
Madrid 28050
Spain

Email: manuel-julian.lopez@vodafone.com

Tianle Yang
China Mobile
32, Xuanwumenxi Ave.
Beijing 01719
China

Email: yangtianle@chinamobile.com

MPLS Working Group
Internet-Draft
Intended status: Standards Track
Expires: January 02, 2014

R. Li
Q. Zhao
Huawei Technologies
C. Jacquenet
France Telecom Orange
E. Metz
KPN
B. Zhang
Telus Communications
July 01, 2013

Receiver-Driven Multicast Traffic-Engineered Label-Switched Paths
draft-lzj-mpls-receiver-driven-multicast-rsvp-te-03.txt

Abstract

This document describes extensions to Resource Reservation Protocol - Traffic Engineering (RSVP-TE) for the setup of Receiver-Driven Traffic-Engineered point-to-multipoint (P2MP) and multipoint-to-multipoint (MP2MP) Label Switched Paths (LSPs) in Multi-Protocol Label Switching (MPLS) and Generalized MPLS (GMPLS) networks.

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 January 02, 2014.

Copyright Notice

Copyright (c) 2013 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
1.1. Motivation	3
1.2. Terminology	4
1.3. Overview	5
2. Receiver-Driven mRSVP-TE LSP Examples	7
2.1. P2MP Example	7
2.2. MP2MP Example	8
3. Signaling Protocol Extensions	9
3.1. Mechanisms	10
3.1.1. Sessions	10
3.1.2. L2S Sub-LSPs	11
3.1.3. Path Originator and Data Receiver	12
3.1.4. Explicit Routing	12
3.2. Path Messages	13
3.3. Resv Messages	14
3.4. PathErr Messages	14
3.5. ResvErr Message	15
3.6. PathTear Messages	15
4. New and Updated Objects	15
4.1. SESSION Objects	15
4.1.1. P2MP LSP for IPv4 SESSION Objects	16
4.1.2. MP2MP LSP for IPv4 SESSION Objects	16
4.1.3. P2MP LSP for IPv6 SESSION Objects	16
4.1.4. MP2MP LSP for IPv6 SESSION Objects	17
4.2. SENDER_TEMPLATE Objects	17
4.2.1. Multicast LSP IPv4 SENDER_TEMPLATE Objects	17
4.2.2. Multicast LSP IPv6 SENDER_TEMPLATE Objects	18

4.3.	L2S_SUB_LSP Objects	18
4.3.1.	L2S_SUB_LSP IPv4 Objects	18
4.3.2.	L2S_SUB_LSP IPv6 Objects	19
4.4.	FILTER_SPEC Objects	19
4.4.1.	mRSVP-TE LSP_IPv4 FILTER_SPEC Objects	19
4.4.2.	mRSVP-TE LSP_IPv6 FILTER_SPEC Objects	19
5.	Applications	20
5.1.	Interwork with PIM	20
5.2.	Multicast VPN	20
6.	Fast Re-Route Considerations	20
7.	Backward Compatibility	21
8.	Acknowledgements	21
9.	IANA Considerations	21
10.	Security Considerations	21
11.	References	22
11.1.	Normative References	22
11.2.	Informative References	23
	Authors' Addresses	24

1. Introduction

Multiparty multimedia applications are getting great attentions in the telecom and datacom world. Such applications are QoS-demanding and can therefore benefit from the MPLS traffic engineering capabilities based on dynamic computation and establishment of MPLS LSPs to meet with application-specific QoS requirements. P2MP-TE [RFC4875] defines a procedure to set up point-to-multipoint LSPs from sender to receivers. This procedure works very well if the senders have a priori knowledge of all its receivers. Sometimes multicast data streams are required to get transported over both IP networks and MPLS networks, but MPLS networks have no priori knowledge about senders and receivers. In the IP networks, the receivers can join/leave a multicast distribution tree by PIM Join/Prune messages, and thus the multicast distribution tree is essentially receiver-driven. When such PIM Join/Prune messages arrive at an MPLS network border, we need a procedure to initiate and set up the multicast distribution tree in MPLS. This document extends RSVP-TE for initiation and setup of P2MP and MP2MP LSPs driven by receivers.

1.1. Motivation

IP multicast distribution trees are initiated by receivers and dynamic by nature. IP multicast applications are also sensitive to bandwidth, especially in the area of residential IPTV services, where the delivery of multicast contents to several hundreds of thousands of IPTV receivers assumes the appropriate level of quality.

Current source-driven P2MP LSP establishment, as defined as in [RFC4875], assumes a priori knowledge of receiver locations, and the LSP signalling is initiated and driven by the data sender (headend). The priori knowledge of receiver locations is obtained either through static configuration or by using another protocol to discover such receivers. On the other hand, [RFC4875] does not address the MP2MP LSPs. Actually, there is no straightforward way to support MP2MP applications by using P2MP LSP unless full-meshed P2MP LSPs are set up independently and separately.

The receiver-driven extension to RSVP-TE described in this document will support both P2MP LSPs and MP2MP LSPs. Moreover, it does not require the sender to know all the receivers' locations a priori. The protocols for discovery of receivers are not needed. It provides a natural mechanism to interwork with PIM dynamically.

1.2. Terminology

The following terms are used in this document:

- o **Sender:** Sender refers to the Originator (and hence the Sender) of the content/payload, as defined in [RFC2205].
- o **Receiver:** Receiver refers to the Receiver of the content/payload, as defined in [RFC2205].
- o **Upstream:** The direction of flow from content Receiver toward content Sender, as defined in [RFC2205].
- o **Downstream:** The direction of flow from content Sender toward content Receiver, as defined in [RFC2205].
- o **Path-Sender:** The sender of RSVP PATH messages, with no correlation to the direction of content/payload flows. Its flow direction is irrelevant to that of Sender defined above. All other control messages discussed in this document will use this as the reference.
- o **Path-Receiver:** The receiver of RSVP PATH messages, with no correlation to the direction of content/payload flows.
- o **Path-Initiator:** The Path-Sender that originated a RSVP PATH message. This is different from Path-Sender in that an intermediate node can be a Path-Sender, but such an intermediate node cannot create and initiate the RSVP PATH message. A Path-Initiator is a Path-Sender, but a Path-Sender doesn't have to be a Path-Initiator.

- o Path-Terminator: The Path-Receiver that does NOT propagate the Path message any further. This is different from Path-Receiver in that an intermediate node can be a Path-Receiver, but such an intermediate node will propagate the Path message to the next hop.
- o Root: A router where a multicast LSP tree is rooted at. Data enters the root and then is distributed to leaves along the P2MP/MP2MP LSP.

1.3. Overview

Although the receiver-driven extensions to RSVP-TE as defined in this document use the existing sender-driven syntax, there are important semantic differences that need to be defined for correct interpretation and interoperability. In the receiver-driven context, we inverted the semantics of RSVP-TE messages, while keeping the syntax unchanged as much as possible. We will use mRSVP-TE to represent the RSVP-TE with receiver-driven extensions described in this document.

The following are some key differences that are specific to the receiver-driven paradigm:

- o The leaf router: the router that receives data/content/payload. In this document, the leaf router will initiate PATH messages. In some sense, the leaf router and the receiver mean the same thing. The term "receiver-driven" also means "leaf-driven".
- o L2S Destinations: routers where user data payload traffic enters the LSP. L2S means Leaf-to-Source. The source is the sender or root of a multicast stream.
- o RSVP P2MP PATH messages traverse from receivers to the root.
- o RSVP P2MP RESV messages traverse from the root to the leaf routers of the P2MP tree structure.
- o For P2MP LSP, a RSVP RESV message received by a router is interpreted as a successful resource reservation made by the upstream node.
- o For MP2MP LSP, a RSVP RESV message received by a router is interpreted as successful resource reservation made by the downstream node.
- o After a PATH message is received on an interface for P2MP LSP, label allocation on that interfaces is done prior to sending the corresponding RSVP PATH message upstream.

- o After a PATH message is received on an interface for MP2MP LSP, label allocation on that interfaces is done prior to sending the corresponding RSVP RESV messages downstream.
- o For P2MP LSP tree structures, a node receiving a RSVP PATH message first decides if this RSVP PATH message will make the said node a branch LSR or not. If it is not a branch LSR, it is a transit LSR. In the case that it will become a transit LSR because of this PATH message, it will, before sending the RSVP PATH message upstream, allocate required bandwidth on the interface on which the RSVP PATH message is received. The upstream node can send traffic soon after successfully reserving resources on the downstream link, on which the RSVP PATH message SHOULD be received. In the case that the node is already a branch or a transit node before it receives the PATH message, then it will allocate required bandwidth on the interface on which the RSVP PATH message is received, and send the RESV message to the node which sends the PATH message without propagating the PATH message further to the upstream node. For P2MP LSPs, a label is carried by the PATH message and should be used by the upstream node when distributing the data from upstream to downstream.
- o For MP2MP LSP tree structures, a node will allocate required bandwidth on the interface through which the RSVP PATH message is sent before sending the RSVP PATH message upstream. A node receiving a RSVP PATH message MUST first decide if this RSVP PATH message will make the said node a branch LSR or not. In the case it will become a transit LSR because of this PATH message, then it will allocate required bandwidth on the interface on which the RSVP PATH message is received and will allocate required bandwidth on the interface through which the RSVP PATH message is sent, before sending the RSVP PATH message upstream. The downstream node can send traffic soon after successfully reserving bandwidth on the upstream link through which the RSVP PATH message SHOULD be sent. The upstream node can send traffic soon after successfully reserving bandwidth on the downstream link on which the RSVP PATH message SHOULD be received. In the case that the node is already a branch or a transit node before it receives the PATH message, then it will allocate required resources on the interface on which the RSVP PATH message is received, and send the RESV message to the node which sends the PATH message without propagating the PATH message further to the upstream node. The label carried by the PATH message should be used by the Path-Receiver node to forward data from the Path-Receiver node to the Path-Sender node, and the label carried by RESV messages should be used by its corresponding Path-Sender node to send data from the Path-Sender node to the Path-Receiver node.

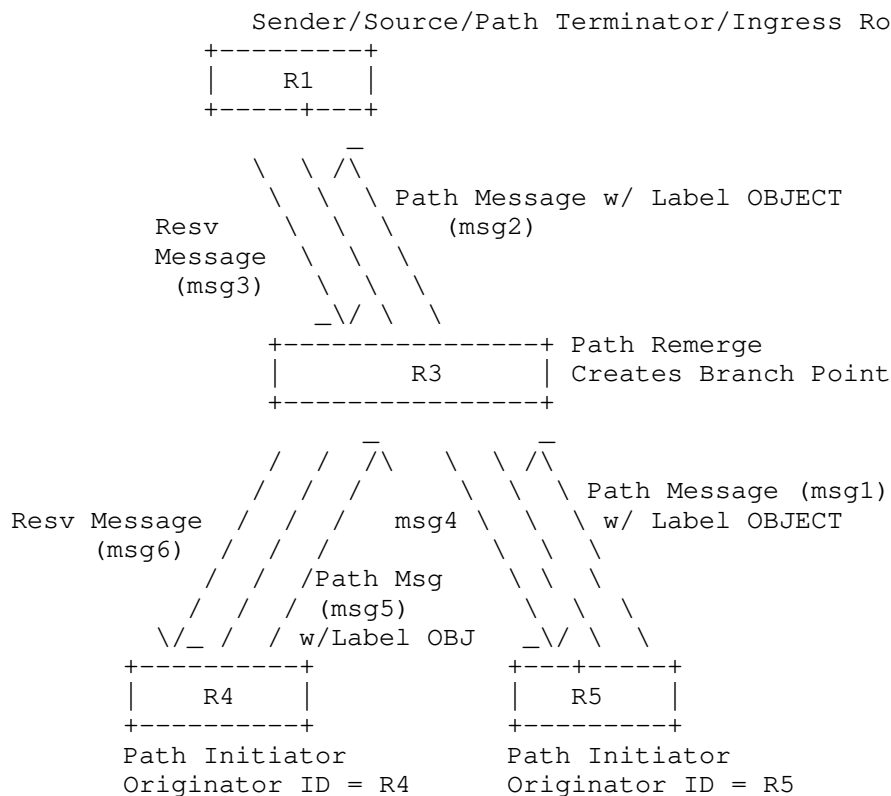
- o For the sake of readability, from now on all mRSVP-TE LSPs will be used to represent all P2MP and/or MP2MP LSPs in receiver-driven (RD) multicast P2MP/MP2MP MPLS environments. We will sometimes use RD P2MP TE LSP or RD MP2MP TE LSP to represent such receiver-driven multicast LSPs.

2. Receiver-Driven mRSVP-TE LSP Examples

In what follows we describe two examples to show how P2MP and MP2MP are set up, respectively. In both of such examples, Path messages are initiated by data receivers.

For the P2MP example, a Path message carries a label for the use of sending data downstream. And for the MP2MP example, both Path message and Resv message carries a label for sending data downstream and upstream.

2.1. P2MP Example

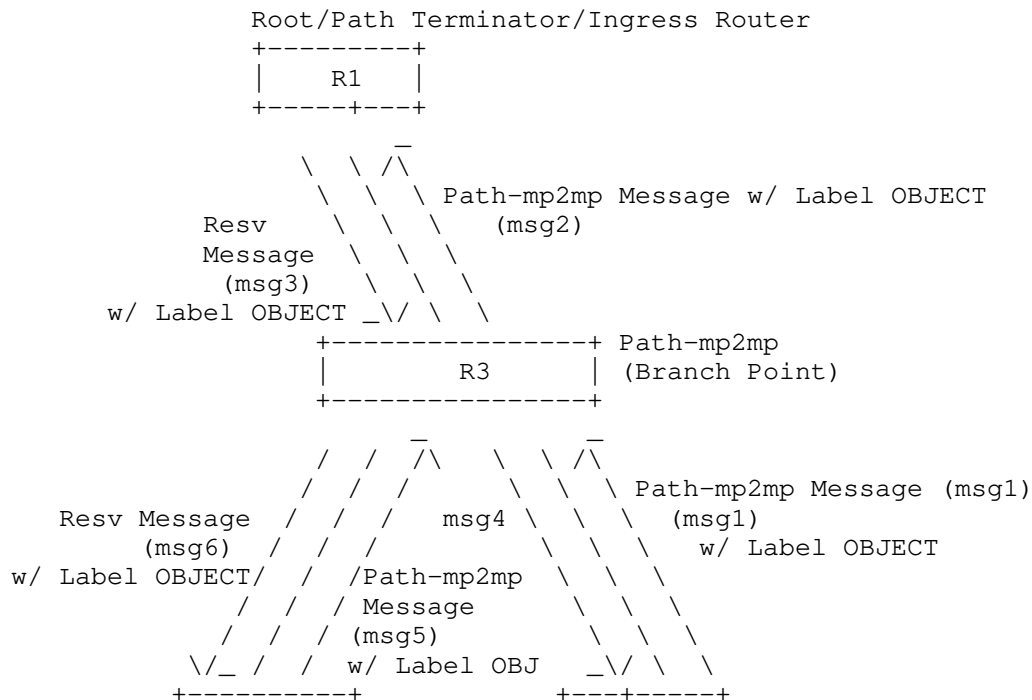


L2S Destination = R1	L2S Destination = R1
Session = S	Session = S

Figure 1: P2MP Example

In Figure 1, when R5 is added as the first leaf of a mulitcast distribution tree (multicast LSP), the message flow goes as follows: R5->msg1->R3->msg2->R1->msg3->R3->msg4->R5. When the leaf R4 is added, the message flow goes from R4->msg5->R3->msg6->R4. In this case, when R3 receives msg5, R3 finds out that a multicast LSP has already been set up for the same session and the same source. Therefore, R3 finds itself a branch node for leaf R4 and R5, so it will terminate the PATH message and build the corresponding RESV message and send it back to R4. The association of the LSP initiated by R4 to the existing multicast LSP is determined based on the processing of the SESSION object and L2S_SUB_LSP object from the mRSVP-TE message. The SESSION object and the L2S_SUB_LSP objects are documented later in this draft.

2.2. MP2MP Example



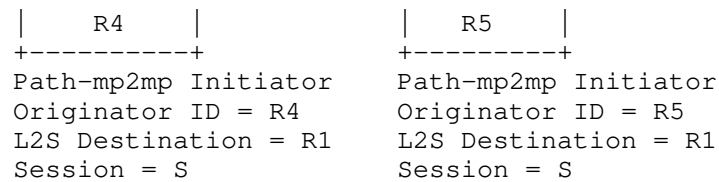


Figure 2: MP2MP Example

For MP2MP, the root address should be specified. It is something similar to RP in PIM, but it doesn't need the Register message. In one-to-many applications, the root should be the same as the Sender, while in many-to-many applications, the root could be any router, but should be selected in the same way as RP is selected in PIM. In Figure 2, R1 is specified as the root. When R5 is added as the first leaf (as both a sender and a receiver) of an MP2MP multicast LSP, the message flow goes from R5->msg1->R3->msg2->R1->msg3->R3->msg4->R5. When the leaf R4 (as both a sender and a receiver) is added, the message flow goes from R4->msg5->R3->msg6->R4. In this case, when R3 receives msg5, R3 finds out that an MP2MP multicast LSP has already been set up for the same session and the same root and R3 will become the branch LSR for the leaf R4 and R5, so it will terminate the PATH message, build a RESV message and send the RESV message back to R4. The association of the LSP initiated by R4 to the existing MP2MP LSP is determined based on the processing of the SESSION object and the S2L_SUB_LSP from the mRSVP-TE message. The SESSION objects and the L2S_SUB_LSP objects are further documented later in this draft.

3. Signaling Protocol Extensions

The RSVP-TE with receiver-driven extensions (mRSVP-TE) is similar to the RSVP-TE protocol as specified in [RFC4875], [RFC3473] and [RFC3209], but differs in that the data receivers of an LSP tunnel initiate the Path messages toward the data sender (or the root of a multicast LSP). Compared with [RFC4875], mRSVP-TE can also be used to set up MP2MP LSPs.

In the context of the receiver-driven RSVP-TE, the Receiver is the Path-Originator. The Path messages go from the Receivers towards the Sender. The Resv messages flow in the opposite direction as compared to the Path messages, i.e. Resv messages are generated by the Sender or a branch LSR. Path messages flow in opposite directions as compared with those of the multicast stream distributions, while Resv messages flow in the same directions as the multicast streams.

In the context of the receiver-driven RSVP-TE, a Path message will be terminated at the "root" of the multicast distribution tree

(multicast LSP) or at an intermediate node if the intermediate node has received another Path message from another receiver for the same multicast distribution tree. When an intermediate node receives two or more Path messages for the same multicast distribution tree, the intermediate node will merge them together. Whether two Path messages should be merged depends on the information encoded in the SESSION and L2S-SUB-LSP objects. The SESSION object encodes multicast group information and the L2S-SUB-LSP (leaf-to-source sub-lsp) object encodes the multicast source or multicast root information.

The following sections describe the receiver-driven extensions to the RSVP-TE protocol. When there is no difference in the protocol, the usage of [RFC4875] is assumed.

3.1. Mechanisms

3.1.1. Sessions

As specified in [RFC2205], a session is a data flow with a particular destination and transport-layer protocol. In the context of multicast, the data flow is essentially a multicast distribution tree rooted at the P2MP source or MP2MP root.

For the sake of reliability, two or more sources/roots may be deployed to distribute the same multicast streams. A multicast stream is often represented by a multicast group address. In this document, we will encode the multicast group address in the SESSION object and the multicast source/root address in the leaf-to-source sub-LSP object. Note that the same session can have different sources/roots, and the same sources/roots can have different sessions.

In the context of the receiver-driven mRSVP-TE, the processing of SESSION objects is different from that of SESSION objects in sender-driven RSVP-TE [RFC4875]. In order to distinguish them, we will employ different C-Types of SESSIONs. In this document we will document SESSION objects for native IPv4/IPv6 multicast applications. For new and more applications, new types of SESSION objects will be added.

Following the method used by RSVP-TE and P2MP RSVP-TE, this draft documents the use of some new SESSION C-Type as follows:

```
Class Name = SESSION
C-Type
  XX+0    mRSVP_TE_P2MP_LSP_TUNNEL_IPv4 C-Type
```

XX+1	mRSVP_TE_P2MP_LSP_TUNNEL_IPv6	C-Type
XX+2	mRSVP_TE_MP2MP_LSP_TUNNEL_IPv4	C-Type
XX+3	mRSVP_TE_MP2MP_LSP_TUNNEL_IPv6	C-Type

Where XX is a number to be allocated by IANA.

Figure 3: New C-Types of SESSIONs

The new SESSION C-Type MUST be used in all receiver-driven P2MP RSVP-TE messages.

3.1.2. L2S Sub-LSPs

A multicast LSP is composed of one or more leaf-to-source sub-LSPs, which are merged together at the branch nodes. There are two ways to identify each such sub-LSP:

- o From the Sender's perspective, each sub-LSP is identified by the SESSION object, the SENDER_TEMPLATE object and S2L_SUB_LSP object, as specified in [RFC 4875]. The SESSION object encodes P2MP ID, Tunnel ID, and Extended Tunnel ID. The P2MP ID is unique within the scope of the sender (ingress LSR) and remains constant throughout the lifetime of the P2MP tree structure. The Extended Tunnel ID, which remains constant throughout the lifetime of the P2MP tree structure, and which should contain the sender's address to make sure the identifier is globally unique. Finally, the Tunnel ID, also remains constant throughout the lifetime of the P2MP tree structure. The SENDER_TEMPLATE object contains the ingress LSR source address. The S2L_SUB_LSP contains the destination address of the sub-LSP.
- o From the Receiver's perspective, each sub-LSP is identified by a new SESSION object, a new SENDER_TEMPLATE object and a new L2S_SUB_LSP object. The SESSION object, different from the one used in typical sender-driven environments, contains information to be used as the key to associate different PATH messages originated from different leaves. The SENDER_TEMPLATE object contains the Path-Originator's address, which is actually the Data Receiver. For P2MP LSP, the L2S_SUB_LSP contains the source address of the sub-LSP, i.e. the data Sender's address. For MP2MP LSP, the L2S_SUB_LSP contains the root address of the sub-LSP. The root address could be any router. The SESSION, SENDER_TEMPLATE and L2S_SUB_LSP all together will identify the multicast stream, the multicast stream's source, and a mulitcast stream's receiver

This document takes the approach from the Receiver's perspective. The approach from the Sender's perspective is documented in [RFC 4875].

Once an LSR receives a receiver-driven Path message with the SESSION object and L2S_SUB_LSP object, the LSR should be able to use the SESSION object and L2S_SUB_LSP object to determine whether the sub-LSP signaled by this Path message should be merged with existing multicast LSPs.

3.1.3. Path Originator and Data Receiver

In the context of the receiver-driven RSVP-TE, a Path Originator is also a Data Receiver. This document will document a new type of SENDER_TEMPLATE object, which contains the Path-Originator's IP address and describes the identity of the Path Originator.

In [RFC 2205] and [RFC 4875], the "sender" is both a path originator and a data sender. In the receiver-driven context, path originators and data senders may be different. For P2MP, path originators are actually the data receivers. For MP2MP, path originators are also both the data senders and data receivers.

In this document, we will use the same Object Class SENDER_TEMPLATE with a different C-Type to represent and identify Path Originator. In the case of P2MP LSP, the SENDER_TEMPLATE describes the identify of a data receiver. In the case of MP2MP, the SENDER_TEMPLATE describes the identify of an LSR which work as both a data sender and a data receiver.

All of the SESSION object, L2S_SUB_LSP object and SENDER_TEMPLATE object together contained in a Path message will uniquely identify a leaf-to-source sub-LSP.

3.1.4. Explicit Routing

An EXPLICIT_ROUTE Object (ERO) is used to optionally specify the explicit route of an L2S sub-LSP. Each signaled ERO corresponds to a particular L2S_SUB_LSP object. Details of explicit route encoding are specified in section 4.5 of [RFC4875], but they are encoded in a reverse order in the receiver-driven context.

When a Path message signals a L2S sub-LSP, the EXPLICIT_ROUTE object encodes the path from the leaf to the root LSR. The Path message also includes the L2S_SUB_LSP object for the L2S sub-LSP being signaled. The < [<EXPLICIT_ROUTE>], <L2S_SUB_LSP>> tuple represents the L2S sub-LSP and is referred to as the sub-LSP descriptor.

The absence of the ERO should be interpreted as requiring hop-by-hop reverse-forwarding for the sub-LSP based on the root address field of the L2S_SUB_LSP object.

3.2. Path Messages

The mechanism specified in this document allows a multicast P2MP/MP2MP LSP to be signaled using one or more Path messages. Each Path message may signal one L2S sub-LSPs.

A receiver-driven P2MP MPLS-TE LSP uses the Path message to carry the LABEL object upstream from the Receiver towards the Sender. With a receiver-driven usage of the RSVP PATH messages, the LABEL_REQUEST object carried by the PATH message is no longer mandatory, it becomes optional for receiver-driven PATH messages, as specified in Figure 4:

```

<Path Message> ::=      <Common Header> [ <INTEGRITY> ]
                        [ [ <MESSAGE_ID_ACK> | <MESSAGE_ID_NACK> ] ... ]
                        [ <MESSAGE_ID> ]
                        <SESSION> <RSVP_HOP>
                        <TIME_VALUES>
                        [ <EXPLICIT_ROUTE> ]
                        [ <LABEL_REQUEST> ]
                        [ <PROTECTION> ]
                        [ <LABEL_SET> ... ]
                        [ <SESSION_ATTRIBUTE> ]
                        [ <NOTIFY_REQUEST> ]
                        [ <ADMIN_STATUS> ]
                        [ <POLICY_DATA> ... ]
                        <sender descriptor>
                        [ <L2S_SUB_LSP> ]

```

Figure 4: Path Message Extensions

The SESSION object encodes information about the being-signalled multicast stream. The SESSION object together with L2S_SUB_LSP will be used as the key to associate different sub-LSPs to the same multicast LSP.

Using [RFC4875] as the base specification, the LABEL object is added to the <sender descriptor> as specified in Figure 5:

```

<sender descriptor> ::= <SENDER_TEMPLATE> <SENDER_TSPEC>

```

```

[ <ADSPEC> ]
[ <RECORD_ROUTE> ]
[ <SUGGESTED_LABEL> ]
[ <RECOVERY_LABEL> ]
<LABEL>

```

Figure 5: Sender Descriptor

The LABEL object is defined in section 4.1 of [RFC3209]

Note that the receiver-driven Path messages convey the LABEL_REQUEST as an optional object. If the Path message signals a P2MP LSP, the LABEL_REQUEST in the Path message is not used. If the Path message signals an MP2MP, the LABEL_REQUEST is needed to ask for labels from its upstream LSR.

3.3. Resv Messages

Receiver-driven P2MP RSVP-TE does not need any change to the basic RESV messages specified in section 6.1 of [RFC4875], as long as the receiver-driven SESSION objects of the new C-Types are used.

For receiver-driven P2MP LSPs, the Path message carries the LABEL object, and thus the Resv message doesn't have to carry the LABEL object anymore. But for MP2MP LSPs, both Path and Resv messages will carry LABEL objects for sending and receiving purposes, respectively. Within the context of MP2MP LSPs, one of the directions is established as per [RFC3209]. Thus, this document is changing the use of the LABEL object in the FF Flow Descriptor and SE Filter Spec from mandatory to optional, as specified in Figure 6:

```

<FF flow descriptor> ::= [ <FLOWSPEC> ] <FILTER_SPEC> [ <LABEL> ]
                        [ <RECORD_ROUTE> ]
                        [ <L2S_SUB_LSP> ]

<SE filter spec> ::=    <FILTER_SPEC> [ <LABEL> ] [ <RECORD_ROUTE> ]
                        [ <L2S_SUB_LSP> ]

```

Figure 6: Resv Message Extensions

3.4. PathErr Messages

The receiver-driven PathErr messages have the same syntax and utilization as the PathErr message described in [RFC4875], with the difference in the <sender descriptor> carried by the PathErr message. The receiver-driven PathErr message will use the <sender descriptor> defined in this document, the same as that carried by the Path messages which the PathErr messages correspond to.

3.5. ResvErr Message

The receiver-driven ResvErr messages have the same syntax and utilization as the ResvErr message described in [RFC4875]. But the ResvErr messages will be processed as per this document, given that the <FF flow descriptor> and the <SE filter spec> can optionally contain the LABEL object instead of mandating the use of the LABEL object. The optional use of the LABEL object is conditioned by the nature of the multicast LSP, either uni-directional (P2MP) or bi-directional (MP2MP).

3.6. PathTear Messages

The receiver-driven PathTear messages have the same syntax and utilization as the PathTear messages described in [RFC4875] except for the <sender descriptor> carried by the PathTear messages. The receiver-driven PathTear messages will use <sender descriptor> defined in this document, the same as that carried by the Path messages which the PathTear messages correspond to.

4. New and Updated Objects

4.1. SESSION Objects

An mRSVP-TE LSP SESSION object is used to represent a multicast stream whose traffic will be carried by the multicast LSP being set up by the mRSVP-TE. The object still uses the existing SESSION C-Num assigned for RSVP-TE, but new C-Types are defined for the new purposes. Different from the values in the existing point-to-point or point-to-multipoint RSVP-TE SESSION object, the new objects defined by the new C-Types will encode "multicasting" information. The new SESSION object will have enough information so that the Path-Receiver can use the SESSION objects together with L2S_SUB_LSP to determine whether or not to associate different Path messages from different leaves to the same P2MP/MP2MP LSP. The combination of the SESSION object, the SENDER_TEMPLATE object and the L2S_SUB_LSP object will uniquely identify a single L2S sub-LSP.

For native IPv4/IPv6 multicast, IPv4/IPv6 (S, G) or (*, G, RP) will be encoded in the SESSION object for P2MP or MP2MP LSPs. In what follows we specify such session objects for IPv4/IPv6 P2MP and MP2MP

applications in the context of receiver-driven RSVP-TE. Other SESSION objects in the receiver-driven context are defined in other documents.

4.1.1. P2MP LSP for IPv4 SESSION Objects

Class = SESSION, mRSVP_TE_P2MP_LSP_TUNNEL_IPv4 C-Type = TBD.

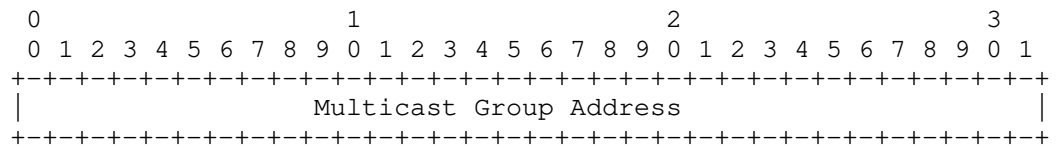


Figure 7: P2MP LSP for IPv4 SESSION Objects

4.1.2. MP2MP LSP for IPv4 SESSION Objects

Class = SESSION, mRSVP_TE_MP2MP_LSP_TUNNEL_IPv4 C-Type = TBD.

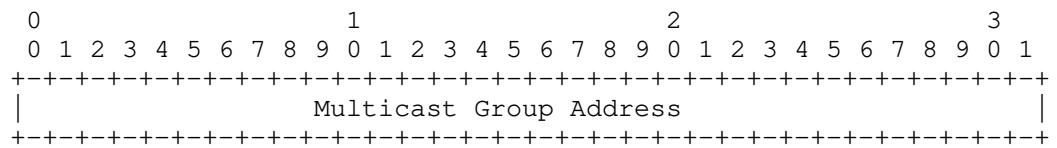


Figure 8: MP2MP LSP for IPv4 SESSION Objects

The MP2MP LSP for IPv4 SESSION objects are of the same format as P2MP LSP for IPv4 SESSION objects, but their C-Types are different.

4.1.3. P2MP LSP for IPv6 SESSION Objects

This is the same as the P2MP LSP for IPv4 SESSION object with the difference that the IPv6 multicast group addresses are 16-byte long.

Class = SESSION, mRSVP_TE_P2MP_LSP_TUNNEL_IPv6 C-Type = TBD.

0 1 2 3

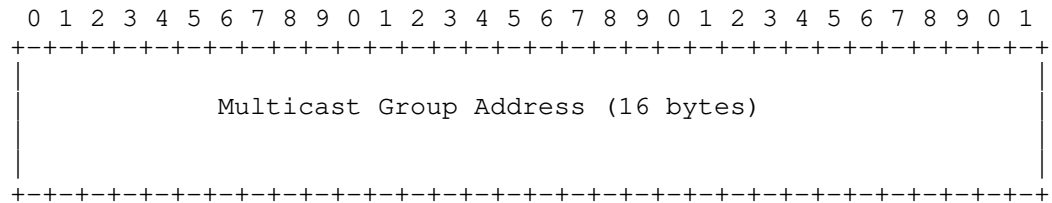


Figure 9: P2MP LSP for IPv6 SESSION Objects

4.1.4. MP2MP LSP for IPv6 SESSION Objects

Class = SESSION, mRSVP_TE_MP2MP_LSP_TUNNEL_IPv6 C-Type = TBD.

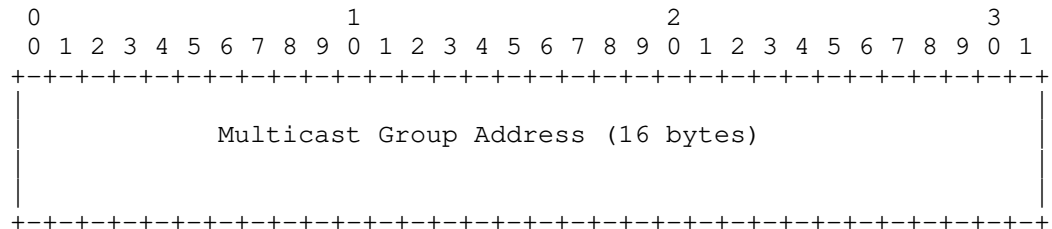


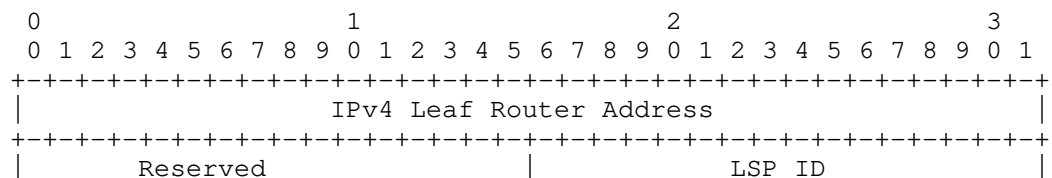
Figure 10: MP2MP LSP for IPv6 SESSION Objects

4.2. SENDER_TEMPLATE Objects

The SENDER_TEMPLATE object contains the Path-Initiator LSR address. In this document, the Path-Initiator is the same as the Leaf Router or Data Receiver. The LSP ID can be changed to allow a sender to do a certain level of resource sharing. Thus, multiple instances of the same mutlicast LSP can be created, each with a different LSP ID. The instances can share resources with each other. The L2S sub-LSPs corresponding to a particular instance use the same LSP ID.

4.2.1. Multicast LSP IPv4 SENDER_TEMPLATE Objects

Class = SENDER_TEMPLATE, mRSVP_TE_LSP_TUNNEL_IPv4 C-Type = TBD.



+ - + - + - + - + - + - + - + - + - + - + - + - + - + - + - + - + - +

Figure 11: mRSVP-TE Multicast LSP SENDER_TEMPLATE Objects

IPv4 Leaf Router Address: The IPv4 address of the Data Receiver.

LSP ID: A 2-byte identifier that can be changed to allow it to share resources with itself. Its usage is the same as that described in [RFC3209].

4.2.2. Multicast LSP IPv6 SENDER_TEMPLATE Objects

Class = SENDER_TEMPLATE, mRSVP-TE_LSP_TUNNEL_IPv6 C-Type = TBD.

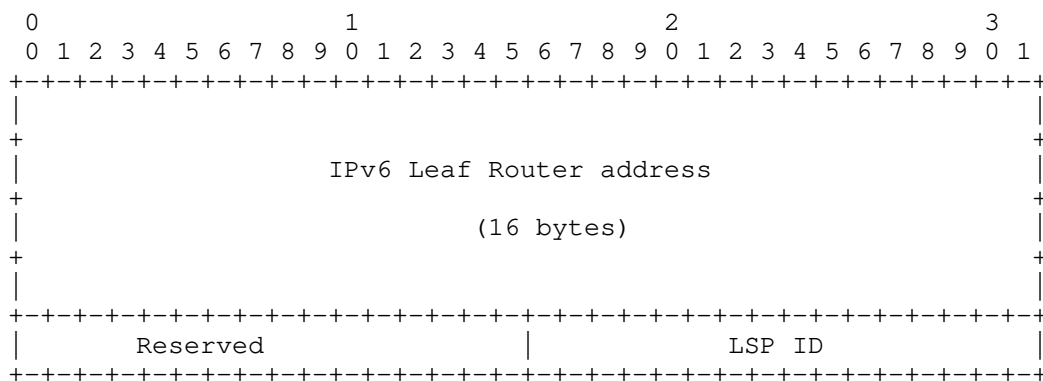


Figure 12: mRSVP-TE LSP IPv6 SENDER_TEMPLATE Objects

IPv6 Leaf Router Address: The IPv6 address of the Data Receiver.

LSP ID: A 2-byte identifier that can be changed to allow it to share resources with itself. Its usage is the same as that described in [RFC3209].

4.3. L2S_SUB_LSP Objects

An `L2S_SUB_LSP` object identifies a particular L2S sub-LSP belonging to a multicast LSP, as explained earlier in this document.

4.3.1. L2S_SUB_LSP IPv4 Objects

L2S_SUB_LSP Class = TBD, L2S_SUB_LSP_IPv4 C-Type = TBD.

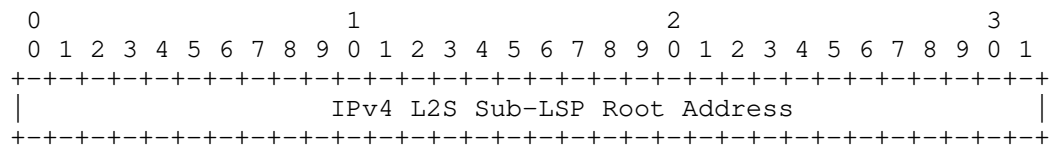


Figure 13: L2S_SUB_LSP IPv4 Objects

IPv4 L2S Sub-LSP Root Address: IPv4 address of the L2S sub-LSP sender.

4.3.2. L2S_SUB_LSP IPv6 Objects

L2S_SUB_LSP Class = TBD, L2S_SUB_LSP_IPv6 C-Type = TBD

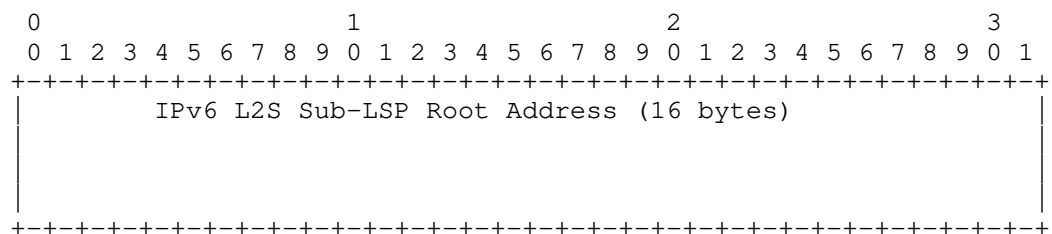


Figure 14: L2S_SUB_LSP IPv6 Object

4.4. FILTER_SPEC Objects

The FILTER_SPEC object is canonical to the SENDER_TEMPLATE object.

4.4.1. mRSVP-TE LSP_IPv4 FILTER_SPEC Objects

Class = FILTER_SPEC, P2MP LSP_IPv4 C-Type = TBD.

The format of the mRSVP-TE LSP_IPv4 FILTER_SPEC object is identical to the mRSVP-TE LSP_TUNNEL_IPv4 SENDER_TEMPLATE object.

4.4.2. mRSVP-TE LSP_IPv6 FILTER_SPEC Objects

The format of the mRSVP-TE LSP_IPv6 FILTER_SPEC object is identical to the mRSVP TE LSP TUNNEL IPv6 SENDER TEMPLATE object.

5. Applications

There are two basic applications for receiver-driven RSVP-TE: interwork with PIM and Multicast VPN.

5.1. Interwork with PIM

Some multicast applications may involve several domains, some of which are operated with PIM while others are enabled with RSVP-TE. This requires the multicast distribution trees to be computed and set up across different domains with PIM and MPLS configured in different domains. When a PIM Join message is received at the border of the MPLS domain, information encoded from the PIM Join message can be encoded as a receiver-driven RSVP-TE Path message which will set up a multicast distribution LSP across the MPLS domain. The root of such a multicast LSP can encode a PIM Join message by using the information encoded in the RSVP-TE Path message. The result of doing so will enable to build a mulitcast distribution tree across both IP and MPLS domains. The multicast tree will consist of a set of IP multicast sub-trees built by PIM and a set of MPLS multicast LSPs built by the receiver-driven RSVP-TE.

5.2. Multicast VPN

An L3VPN service that supports multicast is known as a Multicast VPN, or MVPN for short. An MVPN needs to connect multiple customer sites where some hosts may be senders, may be receivers and may be both senders and receivers. [RFC 6513] specifies protocols and procedures for Multicast in BGP/MPLS IP VPN, and [RFC 6514] describes the BGP encodings and procedures for exchanging the information elements required by Multicast in MPLS/BGP IP VPNs as specified in RFC 6513.

Consider an MVPN with two or more senders. If P2MP RSVP-TE is used to build the multicast distribution tree for multicast in MPLS/BGP IP VPNs, we will need two or more P2MP LSPs, each such P2MP LSP for each sender, which will increase the forwarding states in core routers. The more senders, the more P2MP LSPs, and the more forwarding states. Instead, we can use the extension and the procedure described in this document to set up a single MP2MP LSP no matter how many senders there are. The use of MP2MP will greatly reduce the number of P2MP LSPs and the forwarding states for multicast in BGP/MPLS IP VPNs.

6. Fast Re-Route Considerations

The Fast Re-Route mechanisms and procedures specified in [RFC 4090] will not be applicable to the receiver-driven extension to RSVP-TE described in this document, since their Path/Resv messages are sent in different directions.

Extensions to mRSVP-TE to support Fast Re-Route are described in the document [I-D.zlj-mpls-mpls-mrsvp-te-frr].

7. Backward Compatibility

A receiver-driven P2MP LSP mechanism uses different C-Types than those in the sender-driven P2MP RSVP-TE. If LSRs do not recognize the receiver-driven C-Types, they will not support the receiver-driven extensions described in this document. LSRs that do not support receiver-driven P2MP-TE LSP, send Path Error [TBD] back to the Path Originator.

The complete discussion on the backward compatibility will be provided in the Next version of the document.

8. Acknowledgements

We would like to thank Lin Han, Katherine Zhao, Robert Tao, Lou Berger and Eric Osborne for their comments, questions, and suggestions on our earlier drafts and presentations in IETF meetings.

9. IANA Considerations

This section is TBD.

10. Security Considerations

How a receiver is authenticated is outside the scope of this document. But we will briefly summarize the requirements which are detailed in the requirements draft.

It is a requirement that any mRSVP-TE solution developed to meet some or all of the requirements expressed in this document MUST include mechanisms to enable the secure establishment and management of mRSVP-TE MPLS-TE LSPs. This includes, but is not limited to:

- o A receiver MUST be authenticated before it is allowed to establish mRSVP-TE LSP with its source, in addition to hop-by-hop security issues identified by in RFC 3209 and RFC 4206.
- o mechanisms to ensure that the ingress LSR of a P2MP LSP is identified;
- o mechanisms to ensure that communicating signaling entities can verify each other's identities;
- o mechanisms to ensure that control plane messages are protected against spoofing and tampering;

- o mechanisms to ensure that unauthorized leaves or branches are not added to the mRSVP-TE LSP; and
- o mechanisms to protect signaling messages from snooping.
- o Note that mRSVP-TE signaling mechanisms built on P2P RSVP-TE signaling are likely to inherit all the security techniques and problems associated with RSVP-TE. These problems may be exacerbated in mRSVP-TE situations where security relationships may need to be maintained between an ingress LSR and multiple egress LSRs. Such issues are similar to security issues for IP multicast.
- o It is a requirement that documents offering solutions for P2MP LSPs MUST have detailed security sections.

11. References

11.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC2702] Awduche, D., Malcolm, J., Agogbua, J., O'Dell, M., and J. McManus, "Requirements for Traffic Engineering Over MPLS", RFC 2702, September 1999.
- [RFC3031] Rosen, E., Viswanathan, A., and R. Callon, "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.
- [RFC4461] Yasukawa, S., "Signaling Requirements for Point-to-Multipoint Traffic-Engineered MPLS Label Switched Paths (LSPs)", RFC 4461, April 2006.
- [RFC4875] Aggarwal, R., Papadimitriou, D., and S. Yasukawa, "Extensions to Resource Reservation Protocol - Traffic Engineering (RSVP-TE) for Point-to-Multipoint TE Label Switched Paths (LSPs)", RFC 4875, May 2007.
- [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.

- [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.
- [RFC2205] Braden, B., Zhang, L., Berson, S., Herzog, S., and S. Jamin, "Resource ReSerVation Protocol (RSVP) -- Version 1 Functional Specification", RFC 2205, September 1997.
- [RFC4090] Pan, P., Swallow, G., and A. Atlas, "Fast Reroute Extensions to RSVP-TE for LSP Tunnels", RFC 4090, May 2005.
- [RFC3471] Berger, L., "Generalized Multi-Protocol Label Switching (GMPLS) Signaling Functional Description", RFC 3471, January 2003.

11.2. Informative References

- [I-D.zlj-mppls-mrsvp-te-frr]
Zhao, K., Li, R., and C. Jacquenet, "Fast Reroute Extensions to Receiver-Driven RSVP-TE for Multicast Tunnels", draft-zlj-mppls-mrsvp-te-frr-00 (work in progress), July 2012.
- [RFC3468] Andersson, L. and G. Swallow, "The Multiprotocol Label Switching (MPLS) Working Group decision on MPLS signaling protocols", RFC 3468, February 2003.
- [RFC3473] Berger, L., "Generalized Multi-Protocol Label Switching (GMPLS) Signaling Resource ReserVation Protocol-Traffic Engineering (RSVP-TE) Extensions", RFC 3473, January 2003.
- [RFC3564] Le Faucheur, F. and W. Lai, "Requirements for Support of Differentiated Services-aware MPLS Traffic Engineering", RFC 3564, July 2003.
- [RFC5467] Berger, L., Takacs, A., Caviglia, D., Fedyk, D., and J. Meuric, "GMPLS Asymmetric Bandwidth Bidirectional Label Switched Paths (LSPs)", RFC 5467, March 2009.
- [RFC6513] Rosen, E. and R. Aggarwal, "Multicast in MPLS/BGP IP VPNs", RFC 6513, February 2012.
- [RFC6514] Aggarwal, R., Rosen, E., Morin, T., and Y. Rekhter, "BGP Encodings and Procedures for Multicast in MPLS/BGP IP VPNs", RFC 6514, February 2012.

Authors' Addresses

Renwei Li
Huawei Technologies
2330 Central Expressway
Santa Clara, CA 95050
USA

Email: renwei.li@huawei.com

Quintin Zhao
Huawei Technologies
Boston, MA
USA

Email: quintin.zhao@huawei.com

Christian Jacquenet
France Telecom Orange
4 rue du Clos Courtel
35512 Cesson Sevigne,
France

Email: christian.jacquenet@orange-ftgroup.com

Eduard Metz
KPN
The Netherlands

Email: eduard.metz@kpn.com

Boris Zhang
Telus Communications
200 Consilium PL Floor 15
Toronto, ON M1H 3J3
Canada

Email: Boris.Zhang@telus.com

Internet Engineering Task Force
Internet-Draft
Obsoletes: 3811 (if approved)
Intended status: Standards Track
Expires: December 30, 2013

V. Manral
Hewlett-Packard Corp.
T. Tsou
Huawei Technologies (USA)
W. Liu
Huawei Technologies
F. Fondelli
Ericsson
June 28, 2013

Definitions of Textual Conventions (TCs) for Multiprotocol Label
Switching (MPLS) Management
draft-manral-mpls-rfc3811bis-03

Abstract

This memo defines a Management Information Base (MIB) module which contains Textual Conventions to represent commonly used Multiprotocol Label Switching (MPLS) management information. The intent is that these TEXTUAL CONVENTIONS (TCs) will be imported and used in MPLS related MIB modules that would otherwise define their own representations.

This document obsoletes RFC3811 as it addresses the need to support IPv6 extended TunnelID's by defining a new TC-MplsNewExtendedTunnelID which suggests using IPv4 address of the ingress or egress LSR for the tunnel for an IPv6 network. Changes from RFC3811 and the effect of the new TC to other related documents are summarized in Section 4 and 5, respectively.

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 December 30, 2013.

Copyright Notice

Copyright (c) 2013 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. MPLS Textual Conventions MIB Definitions | 3 |
| 4. Changes from RFC3811 | 17 |
| 5. Effect of the new TC | 17 |
| 6. Contributors | 17 |
| 7. Acknowledgements | 19 |
| 8. Security Considerations | 19 |
| 9. IANA Considerations | 19 |
| 10. References | 20 |
| 10.1. Normative References | 20 |
| 10.2. Informative References | 21 |
| Authors' Addresses | 22 |

1. Introduction

This document defines a MIB module which contains Textual Conventions (TCs) for Multiprotocol Label Switching (MPLS) networks. These Textual Conventions should be imported by MIB modules which manage MPLS networks. The need to support IPv6 extended TunnelID's is addressed by defining a new TC-MplsNewExtendedTunnelID which may represent an IPv4 address of the ingress or egress LSR for the tunnel for an IPv4 network or an IPv6 network.

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].

For an introduction to the concepts of MPLS, see [RFC3031].

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 [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 ([RFC2578], [RFC2579], and [RFC2580]).

3. MPLS Textual Conventions MIB Definitions

```
MPLS-TC-STD-MIB DEFINITIONS ::= BEGIN

IMPORTS

MODULE-IDENTITY,
Unsigned32, Integer32,
transmission          FROM SNMPv2-SMI          -- [RFC2578]

TEXTUAL-CONVENTION
FROM SNMPv2-TC;          -- [RFC2579]

mplsTCStdMIB MODULE-IDENTITY
LAST-UPDATED "200406030000Z" -- June 3, 2004
ORGANIZATION
"IETF Multiprotocol Label Switching (MPLS) Working
Group."
```

CONTACT-INFO

" Vishwas Manral
Hewlett-Packard Corp.

Email comments to the MPLS WG Mailing List at
mpls@uu.net."

DESCRIPTION

"Copyright (C) The Internet Society (2008). The
initial version of this MIB module was published
in Internet Draft. For full legal notices see the RFC
itself or see:
<http://www.ietf.org/copyrights/ianamib.html>

This MIB module defines TEXTUAL-CONVENTIONS
for concepts used in Multiprotocol Label
Switching (MPLS) networks.

Changes from RFC3811 - MplsExtendedTunnelId"

REVISION "200809080000Z" -- 8 September, 2008

DESCRIPTION

"Initial version published as part of Internet Draft. To be
published as RFC XXXX"

-- RFC Ed.: RFC-editor please fill in XXXX
::= { mplsStdMIB 1 }

mplsStdMIB OBJECT IDENTIFIER

::= { transmission 166 }

MplsAtmVcIdentifier ::= TEXTUAL-CONVENTION

DISPLAY-HINT "d"

STATUS current

DESCRIPTION

"A Label Switching Router (LSR) that
creates LDP sessions on ATM interfaces
uses the VCI or VPI/VCI field to hold the
LDP Label.

VCI values MUST NOT be in the 0-31 range.
The values 0 to 31 are reserved for other uses
by the ITU and ATM Forum. The value
of 32 can only be used for the Control VC,
although values greater than 32 could be
configured for the Control VC.

If a value from 0 to 31 is used for a VCI,
the management entity controlling the LDP

subsystem should reject this with an inconsistentValue error. Also, if the value of 32 is used for a VC which is NOT the Control VC, this should result in an inconsistentValue error."

REFERENCE

"MPLS using LDP and ATM VC Switching, RFC3035."
SYNTAX Integer32 (32..65535)

MplsBitRate ::= TEXTUAL-CONVENTION

DISPLAY-HINT "d"

STATUS current

DESCRIPTION

"If the value of this object is greater than zero, then this represents the bandwidth of this MPLS interface (or Label Switched Path) in units of '1,000 bits per second'.

The value, when greater than zero, represents the bandwidth of this MPLS interface (rounded to the nearest 1,000) in units of 1,000 bits per second. If the bandwidth of the MPLS interface is between $((n * 1000) - 500)$ and $((n * 1000) + 499)$, the value of this object is n , such that $n > 0$.

If the value of this object is 0 (zero), this means that the traffic over this MPLS interface is considered to be best effort."

SYNTAX Unsigned32 (0|1..4294967295)

MplsBurstSize ::= TEXTUAL-CONVENTION

DISPLAY-HINT "d"

STATUS current

DESCRIPTION

"The number of octets of MPLS data that the stream may send back-to-back without concern for policing. The value of zero indicates that an implementation does not support Burst Size."

SYNTAX Unsigned32 (0..4294967295)

MplsExtendedTunnelId ::= TEXTUAL-CONVENTION

STATUS obsolete

DESCRIPTION

"A unique identifier for an MPLS Tunnel. This may represent an IPv4 address of the ingress or egress LSR for the tunnel. This value is derived from the Extended Tunnel Id in RSVP or the Ingress Router ID

for CR-LDP."

REFERENCE

"RSVP-TE: Extensions to RSVP for LSP Tunnels,
[RFC3209].

Constraint-Based LSP Setup using LDP, [RFC3212]."

SYNTAX Unsigned32(0..4294967295)

MplsLabel ::= TEXTUAL-CONVENTION

STATUS current

DESCRIPTION

"This value represents an MPLS label as defined in
[RFC3031], [RFC3032], [RFC3034], [RFC3035] and
[RFC3471].

The label contents are specific to the label being
represented, such as:

- * The label carried in an MPLS shim header
(for LDP this is the Generic Label) is a 20-bit
number represented by 4 octets. Bits 0-19 contain
a label or a reserved label value. Bits 20-31
MUST be zero.

The following is quoted directly from [RFC3032].
There are several reserved label values:

- i. A value of 0 represents the
'IPv4 Explicit NULL Label'. This label
value is only legal at the bottom of the
label stack. It indicates that the label
stack must be popped, and the forwarding
of the packet must then be based on the
IPv4 header.
- ii. A value of 1 represents the
'Router Alert Label'. This label value is
legal anywhere in the label stack except at
the bottom. When a received packet
contains this label value at the top of
the label stack, it is delivered to a
local software module for processing.
The actual forwarding of the packet
is determined by the label beneath it
in the stack. However, if the packet is
forwarded further, the Router Alert Label
should be pushed back onto the label stack
before forwarding. The use of this label

is analogous to the use of the 'Router Alert Option' in IP packets [RFC2113]. Since this label cannot occur at the bottom of the stack, it is not associated with a particular network layer protocol.

- iii. A value of 2 represents the 'IPv6 Explicit NULL Label'. This label value is only legal at the bottom of the label stack. It indicates that the label stack must be popped, and the forwarding of the packet must then be based on the IPv6 header.
 - iv. A value of 3 represents the 'Implicit NULL Label'. This is a label that an LSR may assign and distribute, but which never actually appears in the encapsulation. When an LSR would otherwise replace the label at the top of the stack with a new label, but the new label is 'Implicit NULL', the LSR will pop the stack instead of doing the replacement. Although this value may never appear in the encapsulation, it needs to be specified in the Label Distribution Protocol, so a value is reserved.
 - v. Values 4-15 are reserved.
- * The frame relay label can be either 10-bits or 23-bits depending on the DLCI field size and the upper 22-bits or upper 9-bits must be zero, respectively.
 - * For an ATM label the lower 16-bits represents the VCI, the next 12-bits represents the VPI and the remaining bits MUST be zero.
 - * The Generalized-MPLS (GMPLS) label contains a value greater than $2^{24}-1$ and used in GMPLS as defined in [RFC3471]."
- REFERENCE
- "Multiprotocol Label Switching Architecture, [RFC3031].

MPLS Label Stack Encoding, [RFC3032].

Use of Label Switching on Frame Relay Networks,
[RFC3034].

MPLS using LDP and ATM VC Switching, [RFC3035].
Generalized Multiprotocol Label Switching
(GMPLS) Architecture, [RFC3471]."
SYNTAX Unsigned32 (0..4294967295)

MplsLabelDistributionMethod ::= TEXTUAL-CONVENTION
STATUS current
DESCRIPTION
"The label distribution method which is also called
the label advertisement mode [RFC3036].
Each interface on an LSR is configured to operate
in either Downstream Unsolicited or Downstream
on Demand."
REFERENCE
"Multiprotocol Label Switching Architecture,
[RFC3031].

LDP Specification, RFC3036, Section 2.6.3."
SYNTAX INTEGER {
downstreamOnDemand(1),
downstreamUnsolicited(2)
}

MplsLdpIdentifier ::= TEXTUAL-CONVENTION
DISPLAY-HINT "1d.1d.1d.1d:2d"
STATUS current
DESCRIPTION
"The LDP identifier is a six octet
quantity which is used to identify a
Label Switching Router (LSR) label space.

The first four octets identify the LSR and
must be a globally unique value, such as a
32-bit router ID assigned to the LSR, and the
last two octets identify a specific label
space within the LSR."
SYNTAX OCTET STRING (SIZE (6))

MplsLsrIdentifier ::= TEXTUAL-CONVENTION
STATUS current
DESCRIPTION
"The Label Switching Router (LSR) identifier is the
first 4 bytes of the Label Distribution Protocol

```
(LDP) identifier."
SYNTAX OCTET STRING (SIZE (4))
MplsLdpLabelType ::= TEXTUAL-CONVENTION
STATUS current
DESCRIPTION
"The Layer 2 label types which are defined for MPLS
LDP and/or CR-LDP are generic(1), atm(2), or
frameRelay(3)."
```

```
SYNTAX INTEGER {
generic(1),
atm(2),
frameRelay(3)
}
```

```
MplsLSPID ::= TEXTUAL-CONVENTION
STATUS current
DESCRIPTION
"A unique identifier within an MPLS network that is
assigned to each LSP. This is assigned at the head
end of the LSP and can be used by all LSRs
to identify this LSP. This value is piggybacked by
the signaling protocol when this LSP is signaled
within the network. This identifier can then be
used at each LSR to identify which labels are
being swapped to other labels for this LSP. This
object can also be used to disambiguate LSPs that
share the same RSVP sessions between the same
source and destination.
```

For LSPs established using CR-LDP, the LSPID is composed of the ingress LSR Router ID (or any of its own IPv4 addresses) and a locally unique CR-LSP ID to that LSR. The first two bytes carry the CR-LSPID, and the remaining 4 bytes carry the Router ID. The LSPID is useful in network management, in CR-LSP repair, and in using an already established CR-LSP as a hop in an ER-TLV.

For LSPs signaled using RSVP-TE, the LSP ID is defined as a 16-bit (2 byte) identifier used in the SENDER_TEMPLATE and the FILTER_SPEC that can be changed to allow a sender to share resources with itself. The length of this object should only be 2 or 6 bytes. If the length of this octet string is 2 bytes, then it must identify an RSVP-TE LSPID, or it is 6 bytes, it must contain a CR-LDP LSPID."

REFERENCE

"RSVP-TE: Extensions to RSVP for LSP Tunnels,
[RFC3209].

Constraint-Based LSP Setup using LDP,
[RFC3212]."

SYNTAX OCTET STRING (SIZE (2|6))

MplsLspType ::= TEXTUAL-CONVENTION

STATUS current

DESCRIPTION

"Types of Label Switch Paths (LSPs)
on a Label Switching Router (LSR) or a
Label Edge Router (LER) are:

unknown(1) -- if the LSP is not known
to be one of the following.

terminatingLsp(2) -- if the LSP terminates
on the LSR/LER, then this
is an egressing LSP
which ends on the LSR/LER,

originatingLsp(3) -- if the LSP originates
from this LSR/LER, then
this is an ingressing LSP
which is the head-end of
the LSP,

crossConnectingLsp(4) -- if the LSP ingresses
and egresses on the LSR, then it is
cross-connecting on that LSR."

SYNTAX INTEGER {
unknown(1),
terminatingLsp(2),
originatingLsp(3),
crossConnectingLsp(4)
}

MplsOwner ::= TEXTUAL-CONVENTION

STATUS current

DESCRIPTION

"This object indicates the local network
management subsystem that originally created
the object(s) in question. The values of
this enumeration are defined as follows:

unknown(1) - the local network management subsystem cannot discern which component created the object.

other(2) - the local network management subsystem is able to discern which component created the object, but the component is not listed within the following choices, e.g., command line interface (cli).

snmp(3) - The Simple Network Management Protocol was used to configure this object initially.

ldp(4) - The Label Distribution Protocol was used to configure this object initially.

crldp(5) - The Constraint-Based Label Distribution Protocol was used to configure this object initially.

rsvpTe(6) - The Resource Reservation Protocol was used to configure this object initially.

policyAgent(7) - A policy agent (perhaps in combination with one of the above protocols) was used to configure this object initially.

An object created by any of the above choices MAY be modified or destroyed by the same or a different choice."

```
SYNTAX  INTEGER {  
    unknown(1),  
    other(2),  
    snmp(3),  
    ldp(4),  
    crldp(5),  
    rsvpTe(6),  
    policyAgent(7)  
}
```

MplsPathIndexOrZero ::= TEXTUAL-CONVENTION

STATUS current

DESCRIPTION

"A unique identifier used to identify a specific path used by a tunnel. A value of 0 (zero) means that no path is in use."

SYNTAX Unsigned32(0..4294967295)

MplsPathIndex ::= TEXTUAL-CONVENTION
STATUS current
DESCRIPTION
"A unique value to index (by Path number) an entry in a table."
SYNTAX Unsigned32(1..4294967295)

MplsRetentionMode ::= TEXTUAL-CONVENTION
STATUS current
DESCRIPTION
"The label retention mode which specifies whether an LSR maintains a label binding for a FEC learned from a neighbor that is not its next hop for the FEC.

If the value is conservative(1) then advertised label mappings are retained only if they will be used to forward packets, i.e., if label came from a valid next hop.

If the value is liberal(2) then all advertised label mappings are retained whether they are from a valid next hop or not."

REFERENCE
"Multiprotocol Label Switching Architecture, [RFC3031].

LDP Specification, [RFC3036], Section 2.6.2."
SYNTAX INTEGER {
conservative(1),
liberal(2)
}

MplsTunnelAffinity ::= TEXTUAL-CONVENTION
STATUS current
DESCRIPTION
"Describes the configured 32-bit Include-any, include-all, or exclude-all constraint for constraint-based link selection."
REFERENCE
"RSVP-TE: Extensions to RSVP for LSP Tunnels, [RFC3209], Section 4.7.4."
SYNTAX Unsigned32(0..4294967295)

MplsTunnelIndex ::= TEXTUAL-CONVENTION
STATUS current
DESCRIPTION
"A unique index into mplsTunnelTable.

For tunnels signaled using RSVP, this value should correspond to the RSVP Tunnel ID used for the RSVP-TE session."

SYNTAX Unsigned32 (0..65535)

MplsTunnelInstanceIndex ::= TEXTUAL-CONVENTION

STATUS current

DESCRIPTION

"The tunnel entry with instance index 0 should refer to the configured tunnel interface (if one exists).

Values greater than 0, but less than or equal to 65535, should be used to indicate signaled (or backup) tunnel LSP instances. For tunnel LSPs signaled using RSVP, this value should correspond to the RSVP LSP ID used for the RSVP-TE LSP.

Values greater than 65535 apply to FRR detour instances."

SYNTAX Unsigned32 (0|1..65535|65536..4294967295)

TeHopAddressType ::= TEXTUAL-CONVENTION

STATUS current

DESCRIPTION

"A value that represents a type of address for a Traffic Engineered (TE) Tunnel hop.

unknown(0) An unknown address type. This value MUST be used if the value of the corresponding TeHopAddress object is a zero-length string. It may also be used to indicate a TeHopAddress which is not in one of the formats defined below.

ipv4(1) An IPv4 network address as defined by the InetAddressIPv4 TEXTUAL-CONVENTION [RFC3291].

ipv6(2) A global IPv6 address as defined by the InetAddressIPv6 TEXTUAL-CONVENTION [RFC3291].

asnumber(3) An Autonomous System (AS) number as defined by the TeHopAddressAS

TEXTUAL-CONVENTION.

unnum(4) An unnumbered interface index as
defined by the TeHopAddressUnnum
TEXTUAL-CONVENTION.

lspid(5) An LSP ID for TE Tunnels
(RFC3212) as defined by the
MplsLSPID TEXTUAL-CONVENTION.

Each definition of a concrete TeHopAddressType
value must be accompanied by a definition
of a TEXTUAL-CONVENTION for use with that
TeHopAddress.

To support future extensions, the TeHopAddressType
TEXTUAL-CONVENTION SHOULD NOT be sub-typed in
object type definitions. It MAY be sub-typed in
compliance statements in order to require only a
subset of these address types for a compliant
implementation.

Implementations must ensure that TeHopAddressType
objects and any dependent objects
(e.g., TeHopAddress objects) are consistent.
An inconsistentValue error must be generated
if an attempt to change a TeHopAddressType
object would, for example, lead to an
undefined TeHopAddress value that is
not defined herein. In particular,
TeHopAddressType/TeHopAddress pairs
must be changed together if the address
type changes (e.g., from ipv6(2) to ipv4(1))."
REFERENCE
"TEXTUAL-CONVENTIONS for Internet Network
Addresses, [RFC3291].

Constraint-Based LSP Setup using LDP,
[RFC3212]"

SYNTAX INTEGER {
unknown(0),
ipv4(1),
ipv6(2),
asnumber(3),
unnum(4),

```
lspid(5)
}
```

```
TeHopAddress ::= TEXTUAL-CONVENTION
STATUS      current
DESCRIPTION
"Denotes a generic Tunnel hop address,
that is, the address of a node which
an LSP traverses, including the source
and destination nodes. An address may be
very concrete, for example, an IPv4 host
address (i.e., with prefix length 32);
if this IPv4 address is an interface
address, then that particular interface
must be traversed. An address may also
specify an 'abstract node', for example,
an IPv4 address with prefix length
less than 32, in which case, the LSP
can traverse any node whose address
falls in that range. An address may
also specify an Autonomous System (AS),
in which case the LSP can traverse any
node that falls within that AS.
```

A TeHopAddress value is always interpreted within the context of an TeHopAddressType value. Every usage of the TeHopAddress TEXTUAL-CONVENTION is required to specify the TeHopAddressType object which provides the context. It is suggested that the TeHopAddressType object is logically registered before the object(s) which use the TeHopAddress TEXTUAL-CONVENTION if they appear in the same logical row.

The value of a TeHopAddress object must always be consistent with the value of the associated TeHopAddressType object. Attempts to set a TeHopAddress object to a value which is inconsistent with the associated TeHopAddressType must fail with an inconsistentValue error."

```
SYNTAX      OCTET STRING (SIZE (0..32))
```

```
TeHopAddressAS ::= TEXTUAL-CONVENTION
STATUS      current
DESCRIPTION
"Represents a two or four octet AS number.
The AS number is represented in network byte
order (MSB first). A two-octet AS number has
```

the two MSB octets set to zero."

REFERENCE

"Textual Conventions for Internet Network Addresses, [RFC3291]. The InetAutonomousSystemsNumber TEXTUAL-CONVENTION has a SYNTAX of Unsigned32, whereas this TC has a SYNTAX of OCTET STRING (SIZE (4)). Both TCs represent an autonomous system number but use different syntaxes to do so."
SYNTAX OCTET STRING (SIZE (4))

TeHopAddressUnnum ::= TEXTUAL-CONVENTION

STATUS current

DESCRIPTION

"Represents an unnumbered interface:

| octets | contents | encoding |
|--------|----------------------|--------------------|
| 1-4 | unnumbered interface | network-byte order |

The corresponding TeHopAddressType value is unnum(5)."

SYNTAX OCTET STRING (SIZE (4))

MplsNewExtendedTunnelId ::= TEXTUAL-CONVENTION

STATUS current

DESCRIPTION

"A unique identifier for an MPLS Tunnel. This may represent an IPv4 address of the ingress or egress LSR for the tunnel for an IPv4 network. For IPv6 this represents an IPv4 address of the ingress or egress LSR for the tunnel for an IPv6 network. This value is derived from the Extended Tunnel Id in RSVP or the Ingress Router ID for CR-LDP."

REFERENCE

"RSVP-TE: Extensions to RSVP for LSP Tunnels, [RFC3209].

Constraint-Based LSP Setup using LDP, [RFC3212]."

SYNTAX OCTET STRING (SIZE (16))

END

4. Changes from RFC3811

Following is the list of technical changes and other fixes from RFC3811.

Main purpose of this work is to address the need to support IPv6 extended TunnelID's by defining a new TC-MplsNewExtendedTunnelID, resulting in the following changes:

Old MplsExtendedTunnelId status is changed to obsolete.

A new defined MplsNewExtendedTunnelId is added as below.

```
MplsNewExtendedTunnelId ::= TEXTUAL-CONVENTION
    STATUS          current
    DESCRIPTION
        "A unique identifier for an MPLS Tunnel. This may
        represent an IPv4 address of the ingress or egress
        LSR for the tunnel for an IPv4 network. For IPv6
        this represents an IPv4 address of the ingress or
        egress LSR for the tunnel for an IPv6 network.
        This value is derived from the
        Extended Tunnel Id in RSVP or the Ingress Router ID
        for CR-LDP."
    REFERENCE
        "RSVP-TE: Extensions to RSVP for LSP Tunnels,
        [RFC3209].

        Constraint-Based LSP Setup using LDP, [RFC3212]."
    SYNTAX  OCTET STRING(SIZE(16))
```

5. Effect of the new TC

The new TC definition for the MPLS Tunnel will have an effect on the MPLS-TE-MIB and MPLS-TC-STD-MIB. Also the following RFCs which use the MIB may have to be updated to accommodate the changed definition: [RFC3209], [RFC3812], [RFC3813], [RFC3212], [RFC4368], [RFC3814], [RFC3815], and [RFC6639].

6. Contributors

This MIB fixes a small issue with the earlier version of this MIB as defined in RFC3811. The earlier document was created by combining TEXTUAL-CONVENTIONS from current MPLS MIBs and a TE-WG MIB. Co-authors on each of these MIBs contributed to the TEXTUAL-CONVENTIONS contained in this MIB and also contributed greatly to the revisions of this document. These co-authors are:

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

Email: rajiv.papneja@huawei.com

Cheenu Srinivasan
Bloomberg L.P.
499 Park Ave.
New York, NY 10022

Phone: +1-212-893-3682
EMail: cheenu@bloomberg.net

Arun Viswanathan
Force10 Networks, Inc.
1440 McCarthy Blvd
Milpitas, CA 95035

Phone: +1-408-571-3516
EMail: arunv@force10networks.com

Hans Sjostrand
ipUnplugged
P.O. Box 101 60
S-121 28 Stockholm, Sweden

Phone: +46-8-725-5900
EMail: hans@ipunplugged.com

Kireeti Kompella
Juniper Networks
1194 Mathilda Ave
Sunnyvale, CA 94089

Phone: +1-408-745-2000
EMail: kireeti@juniper.net

Thomas D. Nadeau
Cisco Systems, Inc.
BXB300/2/
300 Beaver Brook Road

Boxborough, MA 01719

Phone: +1-978-936-1470

EMail: tnadeau@cisco.com

Joan E. Cucchiara

Marconi Communications, Inc.

900 Chelmsford Street

Lowell, MA 01851

Phone: +1-978-275-7400

EMail: jcucchiara@mindspring.com

7. Acknowledgements

The author would like to thank Adrian Farrel and Thomas Nadeau for thier guidance. The earlier editors and contributors would like to thank Mike MacFadden and Adrian Farrel for their helpful comments on several reviews. Also, a special acknowledgement to Bert Wijnen for his many detailed reviews. Bert's assistance and guidance is greatly appreciated.

8. Security Considerations

This module does not define any management objects. Instead, it defines a set of textual conventions which may be used by other MPLS MIB modules to define management objects.

Meaningful security considerations can only be written in the MIB modules that define management objects. Therefore, this document has no impact on the security of the Internet.

9. IANA Considerations

IANA has made a MIB OID assignment under the transmission branch, that is, assigned the mplsStdMIB under { transmission 166 }. This sub-id is requested because 166 is the ifType for mpls(166) and is available under transmission.

In the future, MPLS related standards track MIB modules should be rooted under the mplsStdMIB subtree. The IANA is requested to manage that namespace. New assignments can only be made via a Standards Action as specified in [RFC2434].

The IANA has also assigned { mplsStdMIB 1 } to the MPLS-TC-STD-MIB specified in this document.

10. References

10.1. Normative References

- [RFC2113] Katz, D., "IP Router Alert Option", RFC 2113, February 1997.
- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC2434] Narten, T. and H. Alvestrand, "Guidelines for Writing an IANA Considerations Section in RFCs", BCP 26, RFC 2434, October 1998.
- [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.
- [RFC3031] Rosen, E., Viswanathan, A., and R. Callon, "Multiprotocol Label Switching Architecture", RFC 3031, January 2001.
- [RFC3032] Rosen, E., Tappan, D., Fedorkow, G., Rekhter, Y., Farinacci, D., Li, T., and A. Conta, "MPLS Label Stack Encoding", RFC 3032, January 2001.
- [RFC3034] Conta, A., Doolan, P., and A. Malis, "Use of Label Switching on Frame Relay Networks Specification", RFC 3034, January 2001.
- [RFC3035] Davie, B., Lawrence, J., McCloghrie, K., Rosen, E., Swallow, G., Rekhter, Y., and P. Doolan, "MPLS using LDP and ATM VC Switching", RFC 3035, January 2001.
- [RFC3036] Andersson, L., Doolan, P., Feldman, N., Fredette, A., and B. Thomas, "LDP Specification", RFC 3036, 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.

- [RFC3212] Jamoussi, B., Andersson, L., Callon, R., Dantu, R., Wu, L., Doolan, P., Worster, T., Feldman, N., Fredette, A., Girish, M., Gray, E., Heinanen, J., Kilty, T., and A. Malis, "Constraint-Based LSP Setup using LDP", RFC 3212, January 2002.
- [RFC3291] Daniele, M., Haberman, B., Routhier, S., and J. Schoenwaelder, "Textual Conventions for Internet Network Addresses", RFC 3291, May 2002.
- [RFC3471] Berger, L., "Generalized Multi-Protocol Label Switching (GMPLS) Signaling Functional Description", RFC 3471, January 2003.

10.2. Informative References

- [RFC3410] Case, J., Mundy, R., Partain, D., and B. Stewart, "Introduction and Applicability Statements for Internet-Standard Management Framework", RFC 3410, December 2002.
- [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 Router (LSR) Management Information Base (MIB)", RFC 3813, June 2004.
- [RFC3814] Nadeau, T., Srinivasan, C., and A. Viswanathan, "Multiprotocol Label Switching (MPLS) Forwarding Equivalence Class To Next Hop Label Forwarding Entry (FEC-To-NHLFE) Management Information Base (MIB)", RFC 3814, June 2004.
- [RFC3815] Cucchiara, J., Sjostrand, H., and J. Luciani, "Definitions of Managed Objects for the Multiprotocol Label Switching (MPLS), Label Distribution Protocol (LDP)", RFC 3815, June 2004.
- [RFC4368] Nadeau, T. and S. Hegde, "Multiprotocol Label Switching (MPLS) Label-Controlled Asynchronous Transfer Mode (ATM) and Frame-Relay Management Interface Definition", RFC 4368, January 2006.

[RFC6639] King, D. and M. Venkatesan, "Multiprotocol Label Switching Transport Profile (MPLS-TP) MIB-Based Management Overview", RFC 6639, June 2012.

Authors' Addresses

Vishwas Manral
Hewlett-Packard Corp.
191111 Pruneridge Ave.
Cupertino, CA 95015
USA

Phone: +1-408-447-1497
Email: vishwas.manral@hp.com

Tina Tsou
Huawei Technologies (USA)
2330 Central Expressway
Santa Clara, CA 95050
USA

Phone: +1 408 330 4424
Email: tina.tsou.zouting@huawei.com

Will (Shucheng) Liu
Huawei Technologies
Bantian, Longgang District
Shenzhen 518129
P.R. China

Email: liushucheng@huawei.com

Francesco Fondelli
Ericsson
via Moruzzi 1
Pisa 56100
Italy

Email: francesco.fondelli@ericsson.com

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: January 13, 2014

E. Osborne
Cisco Systems
July 12, 2013

Extended Administrative Groups in MPLS-TE
draft-osborne-mpls-extended-admin-groups-02

Abstract

This document provides additional administrative groups (sometimes referred to as "link colors") to the IGP extensions for MPLS-TE.

Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [RFC2119].

Status of This Memo

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

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

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

This Internet-Draft will expire on January 13, 2014.

Copyright Notice

Copyright (c) 2013 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 | 2 |
| 1.1. Do we need more than 32 bits? | 2 |
| 2. Extended Administrative Groups sub-TLV | 4 |
| 2.1. Packet Format | 4 |
| 2.2. Admin group numbering | 5 |
| 2.3. Backward compatability | 5 |
| 2.3.1. AG and EAG coexistence | 5 |
| 2.3.2. Desire for unadvertised EAG bits | 5 |
| 3. Attribute filters in RSVP | 6 |
| 3.1. EXTENDED_SESSION_ATTRIBUTE_RA | 6 |
| 3.2. Populating the attribute filter fields | 7 |
| 3.3. Formatting a Path message | 7 |
| 3.4. Interpreting the attribute filter fields | 9 |
| 4. Security Considerations | 9 |
| 5. IANA Considerations | 9 |
| 6. Acknowledgements | 10 |
| 7. Normative References | 10 |
| Author's Address | 10 |

1. Introduction

MPLS-TE advertises 32 administrative groups (commonly referred to as "colors" or "link colors") using the Administrative Group sub-TLV of the Link TLV. This is defined for OSPFv2 [RFC3630], OSPFv3 [RFC5329] and ISIS [RFC5305].

This document adds a sub-TLV to the IGP TE extensions, "Extended Administrative Group". This sub-TLV provides for additional administrative groups (link colors) beyond the current limit of 32.

1.1. Do we need more than 32 bits?

The IGP extensions to support MPLS-TE (RFCs 3630 and 5305) define a link TLV known as Administrative Group (AG) with a limit of 32 AGs per link. This property comes from section 6.2 of RFC 2702 [RFC2702]. RFCs 3630 and 5305 describe the mechanics of the TLV; the actual definition of the field comes from RFC 2702:

"[Administrative Groups] can be used to implement many policies with regard to both traffic and resource oriented performance optimization. Specifically,...[AGs] can be used to:

1. Apply uniform policies to a set of resources that do not need to be in the same topological region.
2. Specify the relative preference of sets of resources for path placement of traffic trunks.
3. Explicitly restrict the placement of traffic trunks to specific subsets of resources.
4. Implement generalized inclusion / exclusion policies.
5. Enforce traffic locality containment policies. That is, policies that seek to contain local traffic within specific topological regions of the network.

Additionally, resource class attributes can be used for identification purposes."

The use of 'Specifically' in RFC2702 is not read as normative; that is, the purpose of the quoted text is not to limit the use of AGs to the six listed policies, they are given as examples. However, the listed policies make good grounds to justify increasing the limit from 32.

Networks have grown over time, and MPLS-TE has grown right along with them. Implementing all six policies with only 32 bits gives the operator only five bits per policy with two bits left over. This can be quite constraining; AGs are a bit mask, so five bits does not mean 32 possible values, it means 5. Running a country-wide or world-wide MPLS-TE network with only five possible values for each case is clearly too constraining.

Even if an operator wishes to use AGs to implement only a single policy it is possible to run out of bit values. One such use case is #5, using AGs to constrain traffic within specific topological regions of the network. A large network may well have far more than 32 geographic regions. One particular operator uses AGs to flag network regions down to the metro scale, e.g. Seattle, San Francisco, Dallas, Chicago, St. Louis, etc. MPLS-TE tunnels are then specified with affinities to include or exclude specific metro regions in their path calculation. It is clear that 32 may not be enough even for a US-based network, nevermind a worldwide network.

The Type of the sub-TLV for OSPF and ISIS is TBD. The Length is the size of the Extended Admin Group (EAG) value in bytes. The EAG may be of any length, but MUST be a multiple of 4 bytes. The only limits on EAG size are those which are imposed by protocol-specific or media-specific constraints (e.g. max packet length).

2.2. Admin group numbering

By convention, the existing Administrative Group TLVs are numbered 0 (LSB) to 31 (MSB). The EAG values are a superset of AG. That is, bits 0-31 in the EAG have the same meaning and MUST have the same values as an AG flooded for the same link.

2.3. Backward compatability

There are two things to consider for backward compatibility with existing AG implementations - how do AG and EAG coexist, and what happens if a node has matching criteria for unadvertised EAG bits?

2.3.1. AG and EAG coexistence

If a node advertises EAG it MAY also advertise AG. If a node advertises both AG and EAG then the first 32 bits of the EAG MUST be identical to the advertised AG. If the AG and EAG advertised for a link differ, the EAG MUST take priority. This allows nodes which do not support EAG to obtain some link color information from the network, but also allow for an eventual migration away from AG. If a node advertises EAG without AG then any receiving node SHOULD alert the network operator to this violation via the appropriate mechanism, e.g. syslog.

2.3.2. Desire for unadvertised EAG bits

The existing AG sub-TLV is optional; thus a node may be configured with a preference to include red or exclude blue, and be faced with a link that is not advertising a value for either blue or red. What does an implementation do in this case? It shouldn't assume that red is set, but it is also arguably incorrect to assume that red is NOT set, as a bit must first exist before it can be set to 0.

Practically speaking this has not been an issue for deployments, as many implementations always advertise the AG bits, often with a default value of 0x00000000. However, this issue may be of more concern once EAGs are added to the network. EAGs may exist on some nodes but not others, and the EAG length may be longer for some links than for others.

Each implementation is free to choose its own method for handling this question. However, to encourage maximum interoperability an implementation SHOULD treat specified but unadvertised EAG bits as if they are set to 0. A node MAY provide other (configurable) strategies for handling this case.

3. Attribute filters in RSVP

In addition to updating the IGP sub-TLV, RSVP needs to be extended to provide the ability to signal desired resource affinities. This section provides that update.

3.1. EXTENDED_SESSION_ATTRIBUTE_RA

This section provides the EXTENDED_SESSION_ATTRIBUTE_RA.

[NOTE: This section reads like EXTENDED_SESSION_ATTRIBUTE_RA is another C-Type of the SESSION_ATTRIBUTE Class. Whether it is implemented like this or whether it ultimately gets specified as a new Class is up for discussion and needs to be resolved prior to publication as an RFC.]

RFC 3209 defines two types of SESSION_ATTRIBUTE, one with resource affinities and one without. The former is C-Type 1 and is referred to in this document as SESSION_ATTRIBUTE_RA. The latter is referred to as SESSION_ATTRIBUTE_NO_RA and is C-Type 7.

The Class and C-Type for EXTENDED_SESSION_ATTRIBUTE_RA are 207 and TBD, respectively. The format of the EXTENDED_SESSION_ATTRIBUTE_RA is:

```

      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
+-----+-----+-----+-----+-----+-----+-----+-----+
|                                     Attribute filter length                                     |
+-----+-----+-----+-----+-----+-----+-----+-----+
|//                                     Exclude-any                                     //|
|                                     +-----+-----+-----+-----+-----+-----+-----+
+-----+-----+-----+-----+-----+-----+-----+-----+
|//                                     Include-any                                     //|
|                                     +-----+-----+-----+-----+-----+-----+-----+
+-----+-----+-----+-----+-----+-----+-----+-----+
|//                                     Include-all                                     //|
|                                     +-----+-----+-----+-----+-----+-----+-----+
+-----+-----+-----+-----+-----+-----+-----+-----+

```

| | | | |
|---|--------------|-------|-------------|
| Setup Prio | Holding Prio | Flags | Name Length |
| Session Name (NULL padded display string) | | | |

The Exclude-any, Include-any and Include-all fields are collectively referred to as the "attribute filter fields". All three attribute filter fields MUST be the same length. All fields in the EXTENDED_SESSION_ATTRIBUTE_RA MUST be interpreted exactly as they are in the SESSION_ATTRIBUTE_RA.

The attribute filter length is the sum of the lengths of the three attribute filter fields, in bytes. If the user wishes to convey 128 bits of information in each of the fields, the total length of the attribute filter fields is $3 \times 128 = 384$ bits. The attribute filter length is thus $384/8 = 48$ bytes. The next 4 bytes of the EXTENDED_SESSION_ATTRIBUTE_RA are fixed - setup priority, holding priority, flags and name length - and the remainder of the object is the Session Name. If the user wishes to convey 128 bits of information each of the three attribute filter fields and provides a 64-byte Session Name then the total length of this object in bytes is $4 + 48 + 4 + 64 = 120$ bytes.

3.2. Populating the attribute filter fields

Each attribute filter field MUST be the same length. As with the EAG sub-TLV, each attribute filter field is a multiple of four bytes in length. The length of each field MUST be at least the minimum length necessary to fully convey the headend's matching criteria, and SHOULD be no longer than that. For example, if the headend wishes to Include-any bits 1 and 17 then all three fields MUST be at least 4 bytes in length and SHOULD be no more than 4 bytes in length. If the headend wishes to Include-any bits 1, 17 and 150 then all three fields MUST be at least 20 bytes (160 bits) in length and SHOULD be no longer than 20 bytes.

3.3. Formatting a Path message

[NOTE: Actual bits and bytes to be sorted out later. For now, this section describes the desired behavior without prescribing specific packet formats. Open questions include - do we need to specify a new Class to hold EXTENDED_SESSION_ATTRIBUTE_RA, or can we reuse C-Type? What's legal? What's least likely to break existing implementations? Once that's decided, we also need a section on how to handle errors such as an invalid combination of resource affinities, etc.)]

In order to provide for backward compatibility, a node MAY signal both `SESSION_ATTRIBUTE_RA` and `EXTENDED_SESSION_ATTRIBUTE_RA`. This allows nodes which understand only `SESSION_ATTRIBUTE_RA` to use it, and nodes which understand `EXTENDED_SESSION_ATTRIBUTE_RA` (and thus also understand `SESSION_ATTRIBUTE_RA`) to use it. If a node signals both `SESSION_ATTRIBUTE_RA` and `EXTENDED_SESSION_ATTRIBUTE_RA`, the first 32 bits of the `EXTENDED_SESSION_ATTRIBUTE_RA` MUST match the `SESSION_ATTRIBUTE_RA`. If they do not match, a node SHOULD alert the operator as to this mismatch, and MUST ignore the `SESSION_ATTRIBUTE_RA` in favor of the `EXTENDED_SESSION_ATTRIBUTE_RA`. This is essentially the same behavior as in section 2.3.1 of this document.

A node MUST NOT signal the combination of (`SESSION_ATTRIBUTE_NO_RA` and `EXTENDED_SESSION_ATTRIBUTE_RA`).

A node MAY signal just `EXTENDED_SESSION_ATTRIBUTE_RA`.

A node MAY signal just `EXTENDED_SESSION_ATTRIBUTE_RA`.

A node MUST NOT signal both `SESSION_ATTRIBUTE_RA` and `SESSION_ATTRIBUTE_NO_RA`.

There are eight combinations of [`SESSION_ATTRIBUTE_NO_RA`, `SESSION_ATTRIBUTE_RA`, and `EXTENDED_SESSION_ATTRIBUTE_RA`], including the combination where none of the three is advertised. Their legality is summarized in the following table, using `SA_NO_RA`, `SA_RA` and `ESA_RA` as abbreviated column headers:

| Valid? | SA_NO_RA | SA_RA | ESA_RA |
|--------|----------|-------|--------|
| Y | x | | |
| Y | | x | |
| Y | | | x |
| Y | | x | x |
| N | x | x | |
| N | x | | x |
| N | x | x | x |
| N | | | |

3.4. Interpreting the attribute filter fields

Since the attribute filter fields are of variable length, it is possible that an RSVP message may indicate more bits than a given node has advertised for a link. It is equally possible that an RSVP message may indicate fewer bits than a given node has advertised for a link. In all cases, the shorter of the two fields (the attribute filter field or the locally configured link admin group) MUST be padded with zeros so that both fields are of equal length.

Specifically, length mismatches are to be handled as follows:

The length of any single attribute filter field is A.

The length of the configured link attribute for a given link is C.

If $C > A$, a node MUST pad the received attribute filter field values with zeros so that $C == A$. A node MUST NOT alter the length of the signalled attribute filter field; the zero padding is only local to a given node.

If $A > C$, a node MUST pad the locally configured link attributes with zeros so that $A == C$. A node SHOULD NOT use this information to alter the length of the EAG sub-TLV that it floods.

[NOTE: rfc3209 is unclear about how the attribute filter fields are to be used. The intent appears to be that any bits set to 1 in any of the three attribute filter fields must be considered a match for filtering purposes, and that any bits set to 0 are not used to match. In other words, there is no way to say "the following bits MUST be zero" for any of the attribute filter fields. A 0 in an attribute filter field says "I do not care what the value of this bit is". I am making this inference largely from the text in Include-any and Include-all which says "A null set...automatically passes". If existing implementations treat these fields differently (e.g. a 0 MUST be matched as a zero) then I'd like to know that so I can get the text in this section right.]

4. Security Considerations

This extension adds no new security considerations.

5. IANA Considerations

This document requests a sub-TLV allocation in both OSPF and ISIS, as well as an RSVP C-Type from Class 207.

6. Acknowledgements

Thanks to Santiago Alvarez, Rohit Gupta, Liem Nguyen, Tarek Saad, and Robert Sawaya for their review and comments.

7. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC2702] Awduche, D., Malcolm, J., Agogbua, J., O'Dell, M., and J. McManus, "Requirements for Traffic Engineering Over MPLS", RFC 2702, September 1999.
- [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.
- [RFC3630] Katz, D., Kompella, K., and D. Yeung, "Traffic Engineering (TE) Extensions to OSPF Version 2", RFC 3630, September 2003.
- [RFC5305] Li, T. and H. Smit, "IS-IS Extensions for Traffic Engineering", RFC 5305, October 2008.
- [RFC5329] Ishiguro, K., Manral, V., Davey, A., and A. Lindem, "Traffic Engineering Extensions to OSPF Version 3", RFC 5329, September 2008.

Author's Address

Eric Osborne
Cisco Systems

Email: eosborne@cisco.com

MPLS Working Group
Internet-Draft
Intended status: Standards Track
Expires: January 02, 2014

R. Li
L. Han
Huawei Technologies
July 01, 2013

Carrying Big Labels in BGP-4
draft-renwei-mpls-bgp-big-label-00.txt

Abstract

When BGP is used to distribute a particular route, it can also be used to distribute an MPLS label which is mapped to that route. In some cases, for example, when L3VPN is used to access and connect to virtual networks in data centers, there may be 16 millions of VPN instances on a router. In order to map MPLS labels to VPN instances, big labels are required. This document specifies the method to carry and distribute such big labels by piggybacking the big label mapping information for an IP route in the BGP Update message that is used to distribute the route itself.

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 January 02, 2014.

Copyright Notice

Copyright (c) 2013 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 | 2 |
| 1.1. Requirement Language | 3 |
| 1.2. Terminology | 3 |
| 2. Motivation | 4 |
| 3. Big Labels | 5 |
| 4. AFI/SAFI | 5 |
| 5. Big Label in NLRI | 6 |
| 6. Capability Announcement | 7 |
| 7. IANA Considerations | 7 |
| 8. Security Considerations | 7 |
| 9. References | 7 |
| 9.1. Normative References | 7 |
| 9.2. Informative References | 8 |
| Authors' Addresses | 8 |

1. Introduction

Network virtualization and server virtualization are being designed and deployed in data center networks, and new data encapsulation methods and protocols are being defined and specified, for example, VXLAN, NVGRE and NVO3. The general idea is to add a new virtual network header so that a physical network can be used to support millions (16M) of virtualized overlaid networks. Network overlay virtualization have placed a new requirement on the access method to such virtualized networks.

BGP/MPLS IP VPNs of [RFC2547] and [RFC4364], provide a market-proven technology and solution for end-to-end IP VPNs. In BGP/MPLS IP VPNs, all the customer sites are connected to the service provider networks

through PE-CE link. It is desirable to extend the BGP/MPLS scheme so that customers can access their virtualized networks hosted in a data center by using BGP/MPLS IP VPNs.

In the data plane of BGP/MPLS IP VPNs, the customer VPN/VRF instances are represented by an MPLS label (VPN label) locally assigned by the PE connecting to CE. Since MPLS labels are 20 bits long, a PE can maximally support 1 million VPNs/VRFs, but PE is required to support 16 millions of virtual networks that are being standardized in VXLAN, NVGRE and NVO3. When BGP/MPLS IP VPNs are extended to access virtualized networks in data centers, [I-D.draft-renwei-l3vpn-big-label] describes several use cases and solutions to use big labels to represent the VPN and maps them to virtual network instances.

The big label information mapped to VPN routes can be carried and distributed by BGP-4 in the framework of [RFC3107]. This document specifies the method to carry and distribute such big labels by piggybacking the big label mapping information for an IP route in the BGP Update message that is used to distribute the route itself.

1.1. Requirement 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].

1.2. Terminology

The following terms are used in this document:

| | |
|----------|--|
| PE | Provider Edge, the provider edge router connected to CE. |
| CE | Customer Edge, the customer edge router connected to PE |
| MPLS LSR | MPLS label switch router |
| IANA | Internet Assigned Numbers Authority |
| AFI | Address Family Identifier |
| SAFI | Subsequent Address Family Identifier |

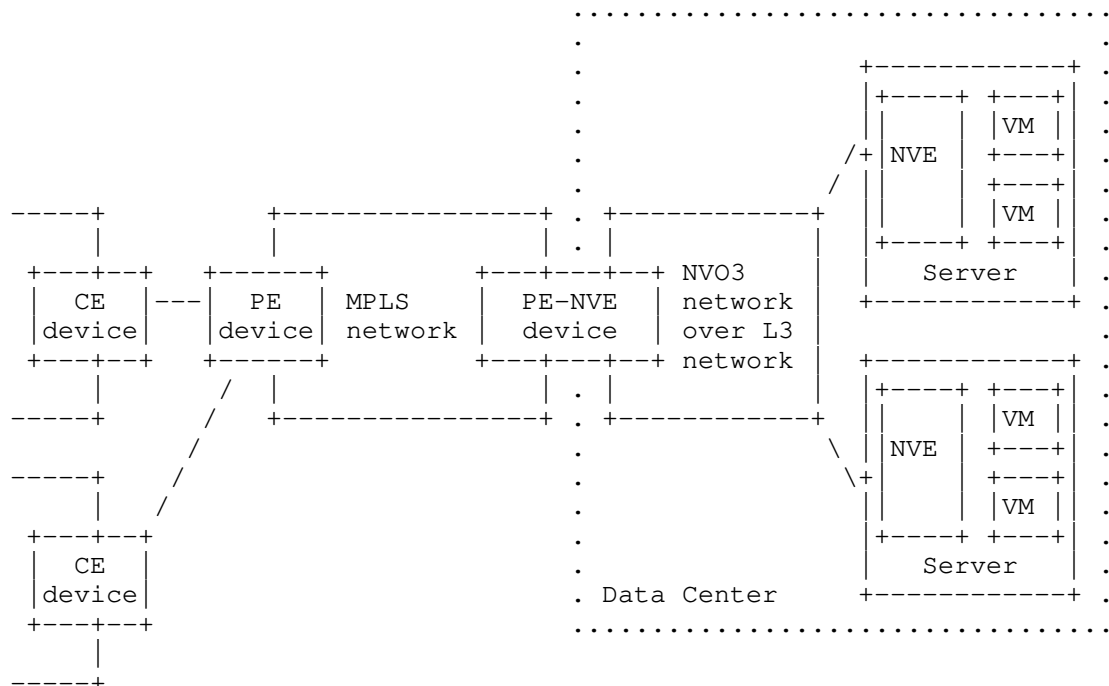
2. Motivation

This document tries to make L3 VPN BGP signaling work with the MPLS big label specified in [I-D.draft-renwei-mpls-big-label]. With the proposed method, the MPLS big label can be carried in BGP NLRI and mapped to BGP routes.

The extension of BGP is used to connect a customer site to its virtualized network hosted in a data center by using, for example, VXLAN, NVGRE or NVO3.

Take NVO3 as an example. NVO3 is an on-going effort to standardize solutions to data center virtualization with the goal of providing viable data encapsulation and protocols across a scaling range of a few thousand VMs to several million VMs running on greater than one hundred thousand physical servers. NVO3 considers approaches to multi-tenancy that reside at the network layer rather than using traditional isolation mechanisms that rely on the underlying layer 2 technology (e.g. VLANs).

Based on NVO3 framework and problem statement, NVO3 will deliver 16 million virtual networks in a physical data center.

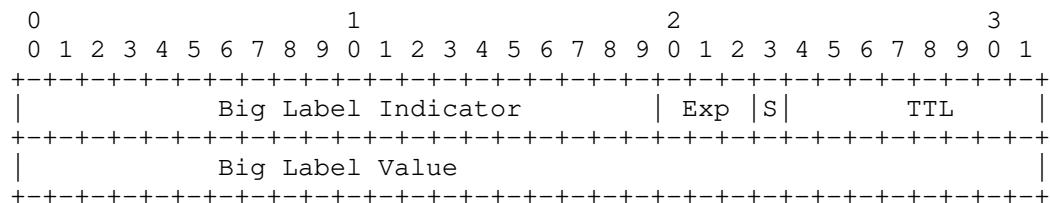


Potentially, PE-NVE needs to support 16 million of VRFs. The extension of BGP described in this document can be used between PE-NVE and PE for the purpose of association between L3VPN labels and network virtualization instances in the NVO3 network.

3. Big Labels

A PE device uses VPN labels to find the associated VRFs for VPN packet forwarding. When L3VPN is used to access and connect a virtual network hosted in a data center by using VXLAN, NVGRE or NVO3, a label space having a minimal set of 16 million labels is required.

The document [I-D.draft-renwei-mpls-big-label] specifies an encoding format by adding a new label value field to the common label as follows:



The Big Label Indicator is a reserved MPLS label. The currently unassigned reserved label range is 4-6 and 8-12. We will temporarily use label 8 for big label indicator, but the final value will be assigned by IANA. The Big Label Value is a 32-bit value.

When an MPLS LSR receives an MPLS packet, it reads out the MPLS label. If the MPLS label is a Big Label Indicator, it will use the subsequent 32-bit value as the MPLS label for the forwarding purpose.

In what follows, we will describe how BGP-4 carries and distributes such big labels for IP routes.

4. AFI/SAFI

In BGP-4, label mapping information is carried as part of the Network Layer Reachability Information (NLRI) in the Multiprotocol Extensions of BGP-4 [RFC2858].

The AFI indicates, as usual, the address family of the associated route. The fact that the NLRI carries an MPLS big label is indicated by using a SAFI value, to be requested to IANA. The SAFI value 4 is currently assigned for common labels. The currently unassigned range is 8-63. Before IANA assigns an official SAFI value for big labels, 8 is temporarily used as the SAFI value to indicate that it carries big labels.

5. Big Label in NLRI

The Network Layer Reachability Information carrying big labels is encoded as one or more triples of the form (length, label, prefix), whose fields are described as follows:

| |
|-------------------|
| Length (1 octet) |
| Label (4 octets) |
| Prefix (variable) |

- a. Length: The Length field indicates the length in bits of the address prefix plus the label.
- b. Label: The Label field carries a 4-octet Big Label Value of the big label format specified in [I-D.draft-renwei-mpls-big-label]. Note that the Big Label Indicator is not carried in NLRI; instead, it will be assigned by IANA. In the data plane, when encoding a packet for forwarding, both the Big Label Indicator and Big Label Value must be encoded in the MPLS header as specified in [I-D.draft-renwei-mpls-big-label].
- c. Prefix: The Prefix field contains address prefixes followed by enough trailing bits to make the end of the field fall on an octet boundary. Note that the value of trailing bits is irrelevant.

All rules and restrictions applicable to the SAFI value 4 is also applicable to the SAFI value 8 (subject to IANA) except for the Label field must be 4 octets. In particular, the following usage rules for SAFI value 4 also applies to the SAFI value for big labels:

The label(s) specified for a particular route (and associated with its address prefix) must be assigned by the LSR which is identified by the value of the Next Hop attribute of the route.

When a BGP speaker redistributes a route, the label(s) assigned to that route must not be changed (except by omission), unless the speaker changes the value of the Next Hop attribute of the route.

A BGP speaker can withdraw a previously advertised route (as well as the binding between this route and a label) by either (a) advertising a new route (and a label) with the same NLRI as the previously advertised route, or (b) listing the NLRI of the previously advertised route in the Withdrawn Routes field of an Update message. The label information carried (as part of NLRI) in the Withdrawn Routes field should be set to 0x800000. (Of course, terminating the BGP session also withdraws all the previously advertised routes.)

6. Capability Announcement

A BGP speaker that uses MP-BGP [RFC2858] to carry label mapping information should use the Capabilities Optional Parameter, as defined in [RFC2842], to inform its peers about this capability. The MP_EXT Capability Code, as defined in [RFC2858], is used to advertise the (AFI, SAFI) pairs available on a particular connection.

A BGP speaker should not advertise this capability to another BGP speaker unless there is a Label Switched Path (LSP) between the two speakers.

7. IANA Considerations

The requirements on IANA are specified in other related documents [I-D.draft-renwei-mpls-big-label] and [I-D.draft-renwei-mpls-bgp-big-label], which request a reserved label to represent Big Label Indicator and BGP capabilities for big labels.

8. Security Considerations

This draft does not add any additional security implications to the BGP/MPLS IP VPNs. All existing authentication and security mechanisms for BGP and MPLS still apply.

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.
- [RFC2547] Rosen, E. and Y. Rekhter, "BGP/MPLS VPNs", RFC 2547, March 1999.

- [RFC4364] Rosen, E. and Y. Rekhter, "BGP/MPLS IP Virtual Private Networks (VPNs)", RFC 4364, February 2006.
- [RFC3107] Rekhter, Y. and E. Rosen, "Carrying Label Information in BGP-4", RFC 3107, May 2001.
- [RFC1771] Rekhter, Y. and T. Li, "A Border Gateway Protocol 4 (BGP-4)", RFC 1771, March 1995.
- [RFC2842] Chandra, R. and J. Scudder, "Capabilities Advertisement with BGP-4", RFC 2842, May 2000.
- [RFC2858] Bates, T., Rekhter, Y., Chandra, R., and D. Katz, "Multiprotocol Extensions for BGP-4", RFC 2858, June 2000.
- [RFC3031] Rosen, E., Viswanathan, A., and R. Callon, "Multiprotocol Label Switching Architecture", RFC 3031, January 2001.
- [RFC3032] Rosen, E., Tappan, D., Fedorkow, G., Rekhter, Y., Farinacci, D., Li, T., and A. Conta, "MPLS Label Stack Encoding", RFC 3032, January 2001.

9.2. Informative References

- [I-D.mahalingam-dutt-dcops-vxlan]
Mahalingam, M., Dutt, D., Duda, K., Agarwal, P., Kreeger, L., Sridhar, T., Bursell, M., and C. Wright, "VXLAN: A Framework for Overlaying Virtualized Layer 2 Networks over Layer 3 Networks", draft-mahalingam-dutt-dcops-vxlan-03 (work in progress), February 2013.
- [I-D.sridharan-virtualization-nvgre]
Sridharan, M., Greenberg, A., Venkataramaiah, N., Wang, Y., Duda, K., Ganga, I., Lin, G., Pearson, M., Thaler, P., and C. Tumuluri, "NVGRE: Network Virtualization using Generic Routing Encapsulation", draft-sridharan-virtualization-nvgre-02 (work in progress), February 2013.

Authors' Addresses

Renwei Li
Huawei Technologies
2330 Central Expressway
Santa Clara, CA 95050
USA

Email: renwei.li@huawei.com

Lin Han
Huawei Technologies
2330 Central Expressway
Santa Clara, CA 95050
USA

Email: lin.han@huawei.com

MPLS Working Group
Internet-Draft
Intended status: Standards Track
Expires: January 01, 2014

R. Li
M. Li
Huawei Technologies
June 30, 2013

Encoding of Big Labels in MPLS Label Stacks
draft-renwei-mpls-big-label-00.txt

Abstract

This document specifies encoding and encapsulation methods for MPLS big labels. Big labels are required for accessing virtual networks in data centers by using, for example, BGP/MPLS IP VPNs. Data center virtualization encapsulation methods and protocols such as VXLAN, NVGRE and NVO3 are being standardized to support a few millions of virtual networks, but the currently label format can support up to one million of labels. When the BGP/MPLS IP VPN method, for example, is used by an enterprise/customer to access its corresponding virtual networks, more than one million of labels are required to map VPN labels and Virtual Network Identifiers.

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 January 01, 2014.

Copyright Notice

Copyright (c) 2013 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 | 2 |
| 1.1. Requirement Language | 3 |
| 1.2. Terminology | 3 |
| 2. Motivations | 3 |
| 3. Review of MPLS Label Stack | 4 |
| 4. Big Labels | 4 |
| 5. IANA Considerations | 5 |
| 6. Security Considerations | 6 |
| 7. References | 6 |
| 7.1. Normative References | 6 |
| 7.2. Informative References | 6 |
| Authors' Addresses | 6 |

1. Introduction

Network virtualization and server virtualization are being designed and deployed in data center networks, and new data encapsulation methods and protocols are being defined and specified, for example, VXLAN, NVGRE and NVO3. The general idea is to add a new virtual network header so that a physical network can be used to support millions (16M) of virtualized overlaid networks. Network overlay virtualization has placed a new requirement on the access method to such huge number of virtualized networks.

BGP/MPLS IP VPNs, as specified in RFC 2547 and RFC 4364, provide a market-proven technology and solution for end-to-end IP VPNs. In BGP/MPLS IP VPNs, all the customer sites are connected to the service provider networks through PE-CE link. It is desirable to extend the

BGP/MPLS scheme so that customers can access their virtualized networks hosted in a data center by using BGP/MPLS IP VPNs.

In the data plane of BGP/MPLS IP VPNs, the customer VPN/VRF instances are represented by an MPLS label (VPN label) locally assigned by the PE connecting to CE. Since MPLS labels are 20 bits long, a PE can maximally support 1 million VPNs/VRFs, but the PE is required to support 16 millions of virtual networks that are being standardized in VXLAN, NVGRE and NVO3. When BGP/MPLS IP VPNs are extended to access virtualized networks in data centers, [I-D.draft-renwei-l3vpn-big-label] specifies use cases and solutions to use big labels to represent the VPN and maps them to virtual network instances.

This document specifies the label format and encoding methods of big labels in the MPLS label stack of [RFC 3032].

1.1. Requirement 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

The following terms are used in this document:

VXLAN

Virtual eXtensible Local Area Network

NVGRE

Network Virtualization using GRE

NVO3

Network Virtualization Overlay over Layer 3

PE

Provider Edge, the provider edge router connected to CE.

CE

Customer Edge, the customer edge router connected to PE

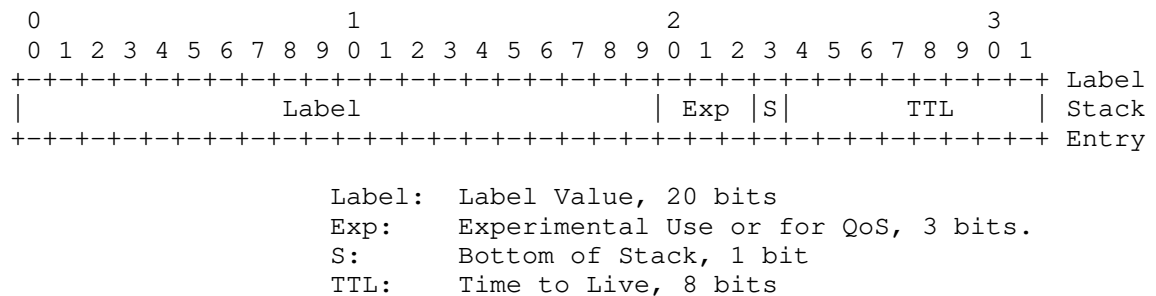
2. Motivations

In [I-D.draft-renwei-l3vpn-big-label], several use cases are described so that an enterprise/customer can use provider-provided BGP/MPLS IP VPN to access its corresponding virtual network hosted in a data center.

The virtual network may be provided by VXLAN, NVGRE or NVO3. In all such network virtualization frameworks, 16 millions of virtual networks may be supported. This implies that up to 16 millions of enterprises/customers can have their own data centers hosted by data center service providers. On the other hand, BGP/MPLS IP VPNs have been used widely by the service providers. This imposes a new requirement of using BGP/MPLS IP VPN protocols and solutions to access the virtual networks in data centers. One problem and obstacle of using BGP/MPLS IP VPN to access virtual networks is that there are not enough labels to do one-one mapping between VPN label space and virtual network identification space.

3. Review of MPLS Label Stack

The label stack is represented as a sequence of "label stack entries". Each label stack entry is represented by 4 octets as follows:



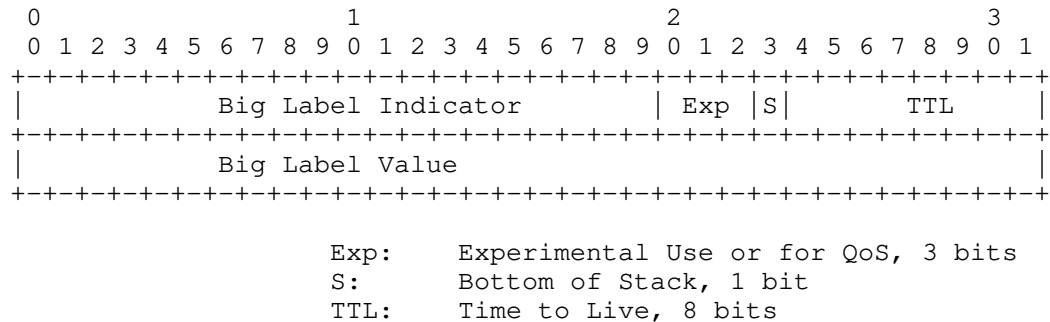
For BGP/MPLS IP VPN, the VPN labels share the same format as all other common MPLS labels as defined as in the above figure.

4. Big Labels

A PE device uses VPN labels to find the associated VRFs for VPN packet forwarding. Since there are potentially 16 millions of virtual networks, 20 bits label are not sufficient; we need to specify a new type of labels: big labels. A big label is an extension to the MPLS label format of RFC 3032 so that the label space is bigger than the 20-bit space with the minimal space being 16 millions of labels.

There are several options to define big labels. One option is to totally re-define the label format; A second option is to extend the length of label entry; A third option is, for the sake of backward compatibility, to add a new field to the common label entry specified in RFC 3032.

The exact format of the third option is defined as follows:



The Big Label Indicator is a reserved MPLS label. The currently unassigned reserved label range is 4-6 and 8-12. We will temporarily use label 8 for big label indicator, but the final value will be assigned by IANA. The Big Label Value is a 32-bit value.

When an MPLS LSR receives an MPLS packet, it reads out the MPLS label. If the MPLS label is a Big Label Indicator, it will use the subsequent 32-bit value as the MPLS label for the forwarding purpose.

All the EXP, S and TTL are also applicable to the Big Label Value as follows:

EXP: Experimental Use or for QoS, 3 bits

S: Bottom of Stack, 1 bit

TTL: Time to Live, 8 bits

5. IANA Considerations

This draft will request IANA to assign a reserved label for Big Label Indicator.

6. Security Considerations

This draft does not add any additional security implications to the BGP/MPLS IP VPNs. All existing authentication and security mechanisms for BGP and MPLS still apply.

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.
- [RFC2547] Rosen, E. and Y. Rekhter, "BGP/MPLS VPNs", RFC 2547, March 1999.
- [RFC3107] Rekhter, Y. and E. Rosen, "Carrying Label Information in BGP-4", RFC 3107, May 2001.
- [RFC3032] Rosen, E., Tappan, D., Fedorkow, G., Rekhter, Y., Farinacci, D., Li, T., and A. Conta, "MPLS Label Stack Encoding", RFC 3032, January 2001.

7.2. Informative References

- [I-D.mahalingam-dutt-dcops-vxlan]
Mahalingam, M., Dutt, D., Duda, K., Agarwal, P., Kreeger, L., Sridhar, T., Bursell, M., and C. Wright, "VXLAN: A Framework for Overlaying Virtualized Layer 2 Networks over Layer 3 Networks", draft-mahalingam-dutt-dcops-vxlan-03 (work in progress), February 2013.
- [I-D.sridharan-virtualization-nvgre]
Sridharan, M., Greenberg, A., Venkataramaiah, N., Wang, Y., Duda, K., Ganga, I., Lin, G., Pearson, M., Thaler, P., and C. Tumuluri, "NVGRE: Network Virtualization using Generic Routing Encapsulation", draft-sridharan-virtualization-nvgre-02 (work in progress), February 2013.

Authors' Addresses

Renwei Li
Huawei Technologies
2330 Central Expressway
Santa Clara, CA 95050
USA

Email: renwei.li@huawei.com

Ming Li
Huawei Technologies
2330 Central Expressway
Santa Clara, CA 95050
USA

Email: mli@huawei.com

MPLS Working Group
Internet-Draft
Intended status: Standards Track
Expires: August 22, 2013

J. Ryoo
ETRI
H. van Helvoort
Huawei Technologies
A. D'Alessandro
Telecom Italia
February 18, 2013

Priority Modification for the PSC Linear Protection
draft-rhd-mpls-tp-psc-priority-00.txt

Abstract

This document contains the modifications to the priorities of inputs in [RFC6378], "MPLS Transport Profile (MPLS-TP) Linear Protection" in an effort to satisfy the ITU-T's protection switching requirements and correcting the problems that have been identified.

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 August 22, 2013.

Copyright Notice

Copyright (c) 2013 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. Motivations for swapping priorities of FS and SF-P | 3 |
| 1.2. Motivation for raising the priority of Clear SF | 4 |
| 2. Conventions Used in This Document | 4 |
| 3. Acronyms | 4 |
| 4. Updates to the PSC RFC | 4 |
| 4.1. Updates to Section 4.3.2. Priority of Inputs | 4 |
| 4.2. Updates to Section 4.3.3.2. Unavailable State | 5 |
| 4.3. Updates to Section 4.3.3.3. Protecting Administrative
State | 5 |
| 4.4. Updates to Appendix A. PSC State Machine Tables | 6 |
| 5. Security considerations | 7 |
| 6. IANA considerations | 7 |
| 7. Acknowledgements | 7 |
| 8. References | 7 |
| 8.1. Normative References | 7 |
| 8.2. Informative References | 8 |
| Appendix A. Freeze Command | 8 |
| Authors' Addresses | 8 |

1. Introduction

This document contains the modifications to the priorities of inputs in [RFC6378], "MPLS Transport Profile (MPLS-TP) Linear Protection" in an effort to satisfy the ITU-T's protection switching requirements and correcting the problems that have been identified.

In this document, the priorities of FS and SF-P are swapped and the priority of Clear SF is raised. The reasons for these changes are explained in the following sub-sections from technical and network operational aspects.

1.1. Motivations for swapping priorities of FS and SF-P

Defining the priority of FS higher than that of SF-P can result in a situation where the protected traffic is taken out-of-service. Setting the priority of any input that is supposed to be signaled to the other end to be higher than that of SF-P can result in unpredictable protection switching state, when the protection path has failed and consequently the PSC communication stopped. An example of the out-of-service scenarios is shown in Annex 1 of the ITU's liaison statement "Liaison Statement: Recommendation ITU-T G.8131/Y.1382 revision - Linear protection switching for MPLS-TP networks" [LIAISON1205].

According to Section 2.4 of [RFC5654] it MUST be possible to operate an MPLS-TP network without using a control plane. This means that external switch commands, e.g. FS, can be transferred to the far end only by using the PSC and should not rely on the presence of a control plane.

As the priority of SF-P has been higher than FS in optical transport networks and Ethernet transport networks, for network operators it is important that the MPLS-TP protection switching preserves the network operation behaviour to which network operators have become accustomed. Typically, the FS command is issued before network maintenance jobs, replacing optical cables or other network components. When an operator pulls out a cable on the protection path by mistake, the traffic should be protected and the operator expects this behaviour based on his/her experience on the traditional transport network operations.

In the case that network operators need an option to control their networks so that the traffic can be placed on the protection path even when the PSC communication channel is broken, an end-to-end command should not be an option. Changing the priority of inputs by provisioning adds complexity and the possibility for mis-configuration. This is unacceptable for transport network operators.

Instead of using FS, the Freeze command, which is a local command and not signaled to the other end, can be used. The use of the Freeze command is described in Appendix A.

1.2. Motivation for raising the priority of Clear SF

The technical issue with the priority of Clear SF defined in [RFC6378] is shown in Appendix IV of [LIAISON1234].

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. Acronyms

This draft uses the following acronyms:

| | |
|---------|----------------------------|
| FS | Forced Switch |
| MPLS-TP | Transport Profile for MPLS |
| SF | Signal Fail |
| SFc | Clear Signal Fail |

4. Updates to the PSC RFC

This section describes the changes required to modify the priorities of FS, SF-P and Clear SF in the PSC protocol defined in [RFC6378]

4.1. Updates to Section 4.3.2. Priority of Inputs

The list of local requests in order of priority should be modified as follows:

- 3 Clear Signal Fail/Degrade (OAM / control-plane / server indication)
- 4 Signal Fail on protection (OAM / control-plane / server indication)
- 5 Forced Switch (operator command)

6 Signal Fail on working (OAM / control-plane / server indication)

7 Signal Degrade on working (OAM / control-plane / server indication)

4.2. Updates to Section 4.3.3.2. Unavailable State

Remove the following bullet items and their text:

- o A local Forced Switch SHALL be ignored by the PSC Control logic when in Unavailable state as a result of a (local or remote) Lockout of protection. If in Unavailable state due to an SF on protection, then the FS SHALL cause the LER to go into local Protecting administrative state and begin transmitting an FS(1,1) message. It should be noted that due to the unavailability of the protection path (i.e., due to the SF condition) that this FS may not be received by the far-end until the SF condition is cleared.
- o A remote Forced Switch message SHALL be ignored by the PSC Control logic when in Unavailable state as a result of a (local or remote) Lockout of protection. If in Unavailable state due to a local or remote SF on protection, then the FS SHALL cause the LER to go into remote Protecting administrative state; if in Unavailable state due to local SF, begin transmitting an SF(0,1) message.

4.3. Updates to Section 4.3.3.3. Protecting Administrative State

Remove the following text in the first paragraph:

The difference between a local FS and local MS affects what local indicators may be received -- the Local Request logic will block any local SF when under the influence of a local FS, whereas the SF would override a local MS.

Replace the following bullet item text:

- o A local Signal Fail indication on the protection path SHALL cause the LER to go into local Unavailable state and begin transmission of an SF(0,0) message, if the current state is due to a (local or remote) Manual Switch operator command. If the LER is in (local or remote) Protecting administrative state due to an FS situation, then the SF on protection SHALL be ignored.

With:

- o A local Signal Fail indication on the protection path SHALL cause the LER to go into local Unavailable state and begin transmission of an SF(0,0) message.

Replace the following bullet item text:

- o A remote Signal Fail message indicating a failure on the protection path SHALL cause the LER to go into remote Unavailable state and begin transmitting an NR(0,0) message, if the Protecting administrative state is due to a Manual Switch command. It should be noted that this automatically cancels the current Manual Switch command and data traffic is reverted to the working path.

With:

- o A remote Signal Fail message indicating a failure on the protection path SHALL cause the LER to go into remote Unavailable state and begin transmitting an NR(0,0) message. It should be noted that this automatically cancels the current Forced Switch or Manual Switch command and data traffic is reverted to the working path.

4.4. Updates to Appendix A. PSC State Machine Tables

Modify the state machine as follows (only modified cells are shown):

Part 1: Local input state machine

| | OC | LO | SF-P | FS | SF-W | SF _c | MS | WTRExp |
|---------|----|----|--------|----|------|-----------------|----|--------|
| N | | | | | | | | |
| UA:LO:L | | | | | | | | |
| UA:P:L | | | | i | | | | |
| UA:LO:R | | | | i | | | | |
| UA:P:R | | | | | | | | |
| PF:W:L | | | | | | | | |
| PF:W:R | | | | | | | | |
| PA:F:L | | | UA:P:L | | | | | |
| PA:M:L | | | | | | | | |
| PA:F:R | | | UA:P:L | | | | | |
| PA:M:R | | | | | | | | |
| WTR | | | | | | | | |
| DNR | | | | | | | | |

Part 2: Remote messages state machine

| | LO | SF-P | FS | SF-W | MS | WTR | DNR | NR |
|---------|----|--------|----|------|----|-----|-----|----|
| N | | | | | | | | |
| UA:LO:L | | | | | | | | |
| UA:P:L | | | i | | | | | |
| UA:LO:R | | | | | | | | |
| UA:P:R | | | i | | | | | |
| PF:W:L | | | | | | | | |
| PF:W:R | | | | | | | | |
| PA:F:L | | UA:P:R | | | | | | |
| PA:M:L | | | | | | | | |
| PA:F:R | | UA:P:R | | | | | | |
| PA:M:R | | | | | | | | |
| WTR | | | | | | | | |
| DNR | | | | | | | | |

Remove the following item in the footnotes for the table:

[19] Transition to PA:F:R and send SF (0,1).

5. Security considerations

No specific security issue is raised in addition to those ones already documented in [RFC6378]

6. IANA considerations

This document makes no request of IANA.

Note to RFC Editor: this section may be removed on publication as an RFC.

7. Acknowledgements

8. References

8.1. Normative References

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

[RFC5654] Niven-Jenkins, B., Brungard, D., Betts, M., Sprecher, N.,

and S. Ueno, "Requirements of an MPLS Transport Profile", RFC 5654, September 2009.

[RFC6378] Weingarten, Y., Bryant, S., Osborne, E., Sprecher, N., and A. Fulignoli, "MPLS Transport Profile (MPLS-TP) Linear Protection", RFC 6378, October 2011.

8.2. Informative References

[LIAISON1205]
ITU-T SG15, "Liaison Statement: Recommendation ITU-T G.8131/Y.1382 revision - Linear protection switching for MPLS-TP networks", <https://datatracker.ietf.org/liaison/1205/> , October 2012.

[LIAISON1234]
ITU-T SG15, "Liaison Statement: Recommendation ITU-T G.8131 revision - Linear protection switching for MPLS-TP networks", <https://datatracker.ietf.org/liaison/1234/> , February 2013.

Appendix A. Freeze Command

The "Freeze" command applies only to the near end (local node) of the protection group and is not signaled to the far end. This command freezes the state of the protection group. Until the Freeze is cleared, additional near end commands are rejected and condition changes and received PSC information are ignored.

"Clear Freeze" command clears the local freeze. When the Freeze command is cleared, the state of the protection group is recomputed based on the persistent condition of the local triggers.

Because the freeze is local, if the freeze is issued at one end only, a failure of protocol can occur as the other end is open to accept any operator command or a fault condition.

Authors' Addresses

Jeong-dong Ryoo
ETRI
218 Gajeongno
Yuseong-gu, Daejeon 305-700
South Korea

Phone: +82-42-860-5384
Email: ryoo@etri.re.kr

Huub van Helvoort
Huawei Technologies

Email: huub.van.helvoort@huawei.com

Alessandro D'Alessandro
Telecom Italia
via Reiss Romoli, 274
Torino 10141
Italy

Phone: +30 011 2285887
Email: alessandro.dalessandro@telecomitalia.it

MPLS Working Group
Internet-Draft
Intended status: Standards Track
Expires: September 13, 2013

J. Ryoo
ETRI
H. van Helvoort
Huawei Technologies
A. D'Alessandro
Telecom Italia
March 12, 2013

Supporting the Signal Degrade in the PSC Linear Protection
draft-rhd-mpls-tp-psc-sd-00.txt

Abstract

This document contains the updates to [RFC6378], "MPLS Transport Profile (MPLS-TP) Linear Protection" to support protection against signal degrade (SD) in an effort to satisfy the ITU-T's protection switching requirements.

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 September 13, 2013.

Copyright Notice

Copyright (c) 2013 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 | 2 |
| 2. Conventions Used in This Document | 3 |
| 3. Acronyms | 3 |
| 4. Operation of Protection Switching against Signal Degrade . . | 3 |
| 5. Updates to the PSC RFC | 4 |
| 5.1. Updates to Section 3.1. Local Request Logic | 4 |
| 5.2. Updates to Section 3.5. Wait-to-Restore (WTR) Timer . . . | 4 |
| 5.3. Updates to Section 3.6. PSC Control States | 5 |
| 5.4. Updates to Section 4.2.2. PSC Request Field | 5 |
| 5.5. Updates to Section 4.2.3. Protection Type (PT) Field . . | 6 |
| 5.6. Updates to Section 4.2.6. Data Path (Path) Field | 6 |
| 5.7. Updates to Section 4.3.2. Priority of Inputs | 6 |
| 5.8. Updates to Section 4.3.3.1 Normal State | 9 |
| 5.9. Updates to Section 4.3.3.2 Unavailable State | 9 |
| 5.10. Updates to Section 4.3.3.3 Protecting Administrative
State | 15 |
| 5.11. Updates to Section 4.3.3.4 Protecting Failure State . . . | 16 |
| 5.12. Updates to Section 4.3.3.5 Wait-to-Restore State | 19 |
| 5.13. Updates to Section 4.3.3.6 Do-not-Revert State | 20 |
| 5.14. Updates to Appendix A. PSC State Machine Tables | 21 |
| 6. Security considerations | 25 |
| 7. IANA considerations | 25 |
| 8. Acknowledgements | 26 |
| 9. References | 26 |
| 9.1. Normative References | 26 |
| 9.2. Informative References | 26 |
| Authors' Addresses | 26 |

1. Introduction

This document contains the updates to [RFC6378], "MPLS Transport Profile (MPLS-TP) Linear Protection" to support protection against signal degrade in an effort to satisfy the ITU-T's protection switching requirements shown in the ITU-T's liaison statements [LIAISON1205] and [LIAISON1234]. In MPLS-TP survivability framework [RFC6372], a fault condition includes both Signal Fail (SF) and Signal Degrade (SD) that can be used to trigger protection switching.

The PSC document [RFC6378] does not specify how the SF and SD are declared but specifies the protection switching protocol associated with SF only.

This document is intended to cover the protection switching protocol associated with SD, and the specifics for the method of identifying SD is out of the scope of this document similarly to SF for [RFC6378]. The updates specified in this document do not require any changes to the protocol's packet format.

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. Acronyms

This draft uses the following acronyms:

| | |
|---------|--------------------------------------|
| FS | Forced Switch |
| LO | Lockout of protection |
| MS | Manual Switch |
| MPLS-TP | Transport Profile for MPLS |
| SD | Signal Degrad |
| SD-P | Signal Degrad on the Protection path |
| SD-W | Signal Degrad on the Working path |
| SF | Signal Fail |
| SF-P | Signal Fail on the Protection path |
| SF-W | Signal Fail on the Working path |
| SFc | Clear Signal Fail |

4. Operation of Protection Switching against Signal Degrad

In order to maintain the network operation behaviour to which transport network operators have become accustomed, the priorities of SD-P and SD-W are defined to be equal as in other transport networks, such as OTN and Ethernet. Once a switch has been completed due to signal degrad on one path, it will not be overridden by signal degrad on the other path (first come, first served behaviour), to avoid protection switching that cannot improve signal quality and flapping.

When multiple SDs are detected simultaneously, either as local or remote requests on both working and protection paths, the SD on the standby path (the path from which the selector does not select the user data traffic) is considered as having higher priority than the SD on the active path (the path from which the selector selects the user data traffic). Therefore, no unnecessary protection switching is performed and the user data traffic continues to be selected from the active path.

In the preceding paragraph, "simultaneously" relates to the occurrence of SD on both the active and standby paths at input to the PSC Protection State Control Logic at the same time, or as long as a SD request has not been acknowledged by the remote end in bidirectional protection switching. In other words, when a local node that has transmitted a SD message receives a SD message that indicates a different value of data path (Path) field than the value of the Path field in the transmitted SD message, both the local and the remote SD requests are considered to occur simultaneously.

5. Updates to the PSC RFC

This section describes the changes required to support protection against SD in the PSC protocol defined in [RFC6378]

5.1. Updates to Section 3.1. Local Request Logic

Replace the following two bullet item text:

- o Signal Degrade (SD) - if any of the server-layer, control-plane, or OAM indications signaled a degraded transmission condition on either the protection path or one of the working paths. The determination and actions for SD are for further study and may appear in a separate document. All references to SD input are placeholders for this extension.
- o Clear Signal Fail (SFc) - if all of the server-layer, controlplane, or OAM indications are no longer indicating a failure condition on a path that was previously indicating a failure condition.

With:

- o Signal Degrade (SD) - if any of the server-layer, control-plane, or OAM indications signaled a degraded transmission condition on either the protection path or one of the working paths.
- o Clear Signal Fail (SFc) - if all of the server-layer, controlplane, or OAM indications are no longer indicating a failure/degradation condition on a path that was previously indicating a failure/degradation condition.

5.2. Updates to Section 3.5. Wait-to-Restore (WTR) Timer

Replace the following text in the first paragraph:

The WTR timer is used to delay reversion to Normal state when recovering from a failure condition on the working path and the protection domain is configured for revertive behavior.

With:

The WTR timer is used to delay reversion to Normal state when recovering from a failure or a degradation condition on the working path and the protection domain is configured for revertive behavior.

5.3. Updates to Section 3.6. PSC Control States

The second paragraph of Section 4.3.3.2 Unavailable State in [RFC6378] shows the intention of including the signal degrade on the protection in the Unavailable state. Even though the protection path can be partially available under the condition of the signal degrade on the protection path, this document follows the same state grouping as [RFC6378] for SD on the protection.

Replace the following bullet item text:

- o Unavailable state - The protection path is unavailable -- either as a result of an operator Lockout command or a failure condition detected on the protection path.

With:

- o Unavailable state - The protection path is unavailable -- either as a result of an operator Lockout command or a failure/degradation condition detected on the protection path.

5.4. Updates to Section 4.2.2. PSC Request Field

Replace the following bullet item text:

- o (7) Signal Degrade - indicates that the transmitting end point has identified a degradation of the signal, or integrity of the packet transmission on either the working or protection path. This request is presented here only as a placeholder. The specifics for the method of identifying this degradation is out of scope for this document. The details of the actions to be taken for this situation are left for future specification.

With:

- o (7) Signal Degrade - indicates that the transmitting end point has identified a degradation of the signal, or integrity of the packet

transmission on either the working or protection path. The FPath field SHALL identify the path that is reporting the degrade condition (i.e., if protection path, then FPath is set to 0; if working path, then FPath is set to 1), and the Path field SHALL indicate where the data traffic is being transported (i.e., if working path is selected, then Path is set to 0; if protection path is selected, then Path is set to 1).

5.5. Updates to Section 4.2.3. Protection Type (PT) Field

Add the following text at the end of Section 4.2.3:

If the detection of a SD depends on the presence of user data packets, such a condition declared on the working path is cleared following protection switching to the protection path if a selector bridge is used, possibly resulting in flapping. To avoid flapping, the selector bridge should duplicate the user data traffic and feed it to both working and protection paths under SD condition.

5.6. Updates to Section 4.2.6. Data Path (Path) Field

Replace the following bullet item text:

- o 0: indicates that the protection path is not transporting user data traffic (in 1:n architecture) or transporting redundant user data traffic (in 1+1 architecture).

With:

- o 0: indicates that the protection path is not transporting user data traffic (in 1:n architecture) or transporting redundant user data traffic (in 1+1 architecture or under SD condition in 1:n architecture when the detection of a SD depends on the presence of user data packets)

5.7. Updates to Section 4.3.2. Priority of Inputs

Replace the following bullet item text:

- o Signal Degrade on working (OAM / control-plane / server indication)

With:

- o Signal Degrade on either working or protection (OAM / control-plane / server indication)

Replace the following two paragraphs:

As was noted above, the Local Request logic SHALL always select the local input indicator with the highest priority as the current local request, i.e., only the highest priority local input will be used to affect the control logic. All local inputs with lower priority than this current local request will be ignored.

The remote message from the far-end LER is assigned a priority just below the similar local input. For example, a remote Forced Switch would have a priority just below a local Forced Switch but above a local Signal Fail on protection input. As mentioned in Section 3.6.1, the state transition is determined by the higher priority input between the highest priority local input and the remote message. This also determines the classification of the state as local or remote. The following subsections detail the transition based on the current state and the higher priority of these two inputs.

With:

As was noted above, the Local Request logic SHALL always select the local input indicator with the highest priority as the current local request, i.e., only the highest priority local input will be used to affect the control logic. All local inputs with lower priority than this current local request will be ignored. For local inputs with same priority, first-come, first-served rule is applied. For example, once SD-P (or SD-W) local input is determined as the highest priority local input, then subsequent SD-W (or SD-P) local input will not be presented to the PSC Control logic as the highest local request.

The remote message from the far-end LER is assigned a priority just below the same local input. For example, a remote Forced Switch would have a priority just below a local Forced Switch but above a local Signal Fail on protection input.

However, if the LER is in a remote state due to a remote message, a subsequent local input having the same priority but requesting different action to the control logic, will be considered as having lower priority than the remote message, and will be ignored. For example, if the LER is in remote Unavailable state due to a remote SD-P, then subsequent local SD-W input will be ignored.

It should be noted that there is a reverse case where one LER receives a local input and the other LER receives, simultaneously, a local input with the same priority but requesting different

action. In this case, each of the two LERs receives a subsequent remote message having the same priority but requesting different action, while the LER is in a local state due to the local input. In this case, a priority must be set for the inputs with the same priority regardless of its origin (local input or remote message). For example, one LER receives SD-P as a local input and the other LER receives SP-W as a local input, simultaneously. In this case, the SD on the standby path (the path from which the selector does not select the user data traffic) is considered as having higher priority than the SD on the active path (the path from which the selector selects the user data traffic) regardless of its origin (local or remote message). Therefore, no unnecessary protection switching is performed and the user data traffic continues to be selected from the active path. Giving the higher priority to the SD on the standby path SHALL also be applied to the Local Request logic when two SDs for different paths happen to be presented to the Local Request logic exactly at the same time.

In order to resolve the equal priority conditions described above, following rules are defined:

- (a) If two local inputs having same priority but requesting different action come to the Local Request logic, then the input coming first SHALL be considered to have a higher priority than the other coming later (first-come, first-served).
- (b) If the LER receives both a local input and a remote message with the same priority and requesting the same action, i.e., the same PSC Request Field and the same FPath value, then the local input SHALL be considered to have a higher priority than the remote message.
- (c) If the LER receives both a local input and a remote message with the same priority but requesting different actions, i.e., the same PSC Request Field but different FPath value, then the first-come, first-served rule SHALL be applied. If the remote message comes first, then the state SHALL be a remote state and subsequent local input is ignored. However, if the local input comes first, the first-come, first-served rule cannot be applied and must be viewed as simultaneous condition. This is because the subsequent remote message will not be an acknowledge of the local input by the far-end node. In this case, the priority SHALL be determined by rules for each simultaneous conditions.
- (d) If the LER receives both SD-P and SD-W request either as local input or remote message and the LER is in a local

state, then the SD on the standby path (the path from which the selector does not select the user data traffic) is considered as having higher priority than the SD on the active path (the path from which the selector selects the user data traffic) regardless of its origin (local or remote message). This rule of giving the higher priority to the SD on the standby path SHALL also be applied to the Local Request logic when two SDs for different paths happen to be presented to the Local Request logic exactly at the same time

As mentioned in Section 3.6.1, the state transition is determined by the higher priority input between the highest priority local input and the remote message. This also determines the classification of the state as local or remote. The following subsections detail the transition based on the current state and the higher priority of these two inputs.

5.8. Updates to Section 4.3.3.1 Normal State

Add the following bullet item text to the transitions in reaction to a local input to the LER:

- o A local Signal Degrade indication on the protection path (SD-P) SHALL cause the LER to go into local Unavailable state and begin transmission of an SD(0,0) message.
- o A local Signal Degrade indication on the working path (SD-W) SHALL cause the LER to go into local Protecting failure state and begin transmission of an SD(1,1) message.

Add the following bullet item text to the transitions in reaction to a remote message:

- o A remote SD-P message SHALL cause the LER (LER-A) to go into remote Unavailable state, while continuing to transmit the NR(0,0) message.
- o A remote SD-W message SHALL cause the LER to go into remote Protecting failure state, and transmit an NR(0,1) message.

5.9. Updates to Section 4.3.3.2 Unavailable State

The second paragraph of Section 4.3.3.2 Unavailable State in [RFC6378] shows the intention of including the signal degrade on the protection in the Unavailable state. This document follows the same state grouping as [RFC6378] for SD-P, even though the protection path can be partially available under the condition of the signal degrade on the protection path.

Replace the following text in the first paragraph of Section 4.3.3.2 Unavailable State for further clarification on SD on the protection path:

When the protection path is unavailable -- either as a result of a Lockout operator command, or as a result of a SF detected on the protection path -- then the protection domain is in the Unavailable state.

With:

When the protection path is unavailable -- either as a result of a Lockout operator command, or as a result of a SF/SD detected on the protection path -- then the protection domain is in the Unavailable state.

When an LER is in this state due to degradation condition, the user traffic should be duplicated and fed to both working and protection paths if the detection of a SD depends on the presence of user data packets.

Replace the following bullet item text in the transitions in reaction to a local input:

- o A local Forced Switch SHALL be ignored by the PSC Control logic when in Unavailable state as a result of a (local or remote) Lockout of protection. If in Unavailable state due to an SF on protection, then the FS SHALL cause the LER to go into local Protecting administrative state and begin transmitting an FS(1,1) message. It should be noted that due to the unavailability of the protection path (i.e., due to the SF condition) that this FS may not be received by the far-end until the SF condition is cleared.

With:

- o A local Forced Switch SHALL be ignored by the PSC Control logic when in Unavailable state as a result of a (local or remote) Lockout of protection. If in Unavailable state due to an SF/SD on protection, then the FS SHALL cause the LER to go into local Protecting administrative state and begin transmitting an FS(1,1) message. It should be noted that due to the unavailability of the

protection path (i.e., due to the SF condition) that this FS may not be received by the far-end until the SF condition is cleared.

Replace the following bullet item text in the transitions in reaction to a local input:

- o A local Signal Fail on the protection path input when in local Unavailable state (by implication, this is due to a local SF on protection) SHALL cause the LER to remain in local Unavailable state and transmit an SF(0,0) message.

With:

- o A local Signal Fail on the protection path input when in local Unavailable state SHALL cause the LER to remain in local Unavailable state and transmit an SF(0,0) message.

Replace the following bullet item text in the transitions in reaction to a local input:

- o A local Signal Fail on the working path input when in remote Unavailable state SHALL cause the LER to remain in remote Unavailable state and transmit an SF(1,0) message.

With:

- o A local Signal Fail on the working path input when in local or remote Unavailable state due to SD-P SHALL cause the LER to go to local Protecting failure state. If the LER is in remote Unavailable state due to SF-P or Lockout of protection, the LER SHALL remain in remote Unavailable state and transmit an SF(1,0) message.

Add the following bullet item text to the transitions in reaction to a local input:

- o A local Clear SD of the protection path in local Unavailable state that is due to an SD on the protection path SHALL cause the LER to go to Normal state. If the LER is in remote Unavailable state but has an active local SD condition, then the local Clear SD SHALL clear the SD local condition and the LER SHALL remain in remote Unavailable state and begin transmitting NR(0,0) messages. In all other cases, the local Clear SD SHALL be ignored.
- o A local SD-P input when in local Unavailable state (by implication, this is due to a local SD on protection) SHALL cause the LER to remain in local Unavailable state and transmit an SD(0,0) message. When in remote Unavailable state due to LO or

SF-P, the LER SHALL remain in remote unavailable state and begin transmitting SD(0,0) messages. When in remote Unavailable state due to SD-P, the LER SHALL enter to local Unavailable state and begin transmitting SD(0,0) messages.

- o A local SD-W input when in remote Unavailable state SHALL cause the LER to remain in remote Unavailable state and transmit an SD(1,0) message.

Replace the following bullet item text in the transitions in reaction to a remote message:

- o A remote Lockout of protection message SHALL cause the LER to remain in Unavailable state (note that if the LER was previously in local Unavailable state due to a Signal Fail on the protection path, then it will now be in remote Unavailable state) and continue transmission of the current message (either NR(0,0) or LO(0,0) or SF(0,0)).

With:

- o A remote Lockout of protection message SHALL cause the LER to remain in Unavailable state (note that if the LER was previously in local Unavailable state due to a Signal Fail on the protection path or a Signal Degrade on the protection path, then it will now be in remote Unavailable state) and continue transmission of the current message (either NR(0,0) or LO(0,0) or SF(0,0) or SF(1,0) or SD(0,0) or SD(1,0)).

Replace the following bullet item text in the transitions in reaction to a remote message:

- o A remote Forced Switch message SHALL be ignored by the PSC Control logic when in Unavailable state as a result of a (local or remote) Lockout of protection. If in Unavailable state due to a local or remote SF on protection, then the FS SHALL cause the LER to go into remote Protecting administrative state; if in Unavailable state due to local SF, begin transmitting an SF(0,1) message.

With:

- o A remote Forced Switch message SHALL be ignored by the PSC Control logic when in Unavailable state as a result of a local Lockout of protection. If in Unavailable state due to a remote Lockout of protection, the LER SHALL go to remote Protecting Administrative state and begin transmitting a message reflecting its local input with Path=1. If in Unavailable state due to a local or remote SF-P/SD-P, then the FS SHALL cause the LER to go into remote

Protecting administrative state; if in Unavailable state due to local SF-P and SD-P, begin transmitting an SF(0,1) and SD(0,1) message, respectively.

Replace the following bullet item text in the transitions in reaction to a remote message:

- o A remote Signal Fail message that indicates that the failure is on the protection path SHALL cause the LER to remain in Unavailable state and continue transmission of the current message (either NR(0,0) or SF(0,0) or LO(0,0)).

With:

- o A remote Signal Fail message that indicates that the failure is on the protection path SHALL cause the LER to remain in Unavailable state and continue transmission of the current message (either NR(0,0) or LO(0,0) or SF(0,0) or SF(1,0) or SD(0,0) or SD(1,0)

Replace the following bullet item text in the transitions in reaction to a remote message:

- o A remote No Request, when the LER is in remote Unavailable state and there is no active local Signal Fail SHALL cause the LER to go into Normal state and continue transmission of the current message. If there is a local Signal Fail on the protection path, the LER SHALL remain in local Unavailable state and transmit an SF(0,0) message. If there is a local Signal Fail on the working path, the LER SHALL go into local Protecting Failure state and transmit an SF(1,1) message. When in local Unavailable state, the remote message SHALL be ignored.

With:

- o A remote No Request, when the LER is in remote Unavailable state and there is no active local Signal Fail or Signal Degrade SHALL cause the LER to go into Normal state and continue transmission of the current message. If there is a local Signal Fail on the protection path, the LER SHALL remain in local Unavailable state and transmit an SF(0,0) message. If there is a local Signal Fail on the working path, the LER SHALL go into local Protecting Failure state and transmit an SF(1,1) message. If there is a local Signal Degrade on the protection path, the LER SHALL remain in local Unavailable state and transmit an SD(0,0) message. If there is a local Signal Degrade on the working path, the LER SHALL go into local Protecting Failure state and transmit an SD(1,1) message. When in local Unavailable state, the remote message SHALL be ignored.

Add the following bullet item text to the transitions in reaction to a remote message:

- o A remote SF-W message SHALL be ignored if the LER is in local Unavailable state due to LO or SF-P. When in local Unavailable state due to SD-P, the LER SHALL enter to remote Protecting Failure state and begin transmitting SD(0,1) messages. If the LER is in remote Unavailable state, then the SF-W message and the local input are reevaluated as if the LER is in the Normal state. In the case that the LER is in remote Unavailable state due to remote SD-P, the reevaluation will cause the LER to enter remote Protecting Failure state and continue to send the current messages with Path=1.
- o A remote MS message SHALL be ignored if the LER is in local Unavailable state. If the LER is in remote Unavailable state, then the MS message and the local input are reevaluated as if the LER is in the Normal state.
- o A remote SD-P message shall be ignored if the LER is in local Unavailable state. If the LER is in remote Unavailable state due to LO or SF-P, then the SD-P message and the local input are reevaluated as if the LER is in the Normal state. If the LER is in remote Unavailable state due to SD-P, then the remote SD-P message will be ignored
- o A remote SD-W message shall be reevaluated with the local input as if the LER is in the Normal state, A remote SD-W message shall be ignored if the LER is in local Unavailable state due to LO or SF-P. When in local Unavailable state due to SD-P, the LER shall examine the Path value in the remote SD-W message. If the Path value of the received SD-W message is the same as the Path value that the LER indicates in its current outgoing PSC message, then the LER shall ignore the SD-W message. Otherwise, as the local SD-P and the remote SD-W are considered to occur simultaneously, perform the followings:
 - * If the working path was the active path at the time when local SD-P was selected as the highest local request, the LER remains in the local Unavailable state and continue transmission of the current message.
 - * If the working path was the standby path at the time when local SD-P was selected as the highest local request, the LER enters into the remote Protection Failure state and begin transmitting SD(0,1) messages.

5.10. Updates to Section 4.3.3.3 Protecting Administrative State

Add the following bullet item text to the transitions in reaction to a local input:

- o A local SD-P SHALL cause the LER to go to local Unavailable state and begin transmitting an SD(0,0) message, if the current state is due to a (local or remote) MS command. If the LER is in remote Protecting administrative state due to a remote Forced Switch command, then this local indication SHALL cause the LER to remain in remote Protecting administrative state and transmit an SD(0,1) message. If the LER is in local Protecting administrative state due to a local FS command, then this indication SHALL be ignored (i.e., the indication should have been blocked by the Local Request logic).
- o A local SD-W SHALL cause the LER to go to local Unavailable state and begin transmitting an SD(1,1) message, if the current state is due to a (local or remote) MS command. If the LER is in remote Protecting administrative state due to a remote Forced Switch command, then this local indication SHALL cause the LER to remain in remote Protecting administrative state and transmit an SD(1,1) message. If the LER is in local Protecting administrative state due to a local FS command, then this indication SHALL be ignored (i.e., the indication should have been blocked by the Local Request logic).

Add the following bullet item text to the transitions in reaction to a remote message:

- o A remote SD-P SHALL cause the LER to go into remote Unavailable state and begin transmitting an NR(0,0) message, if the Protecting administrative state is due to a (local or remote) MS command. It should be noted that this automatically cancels the current MS command and data traffic is reverted to the working path. If the LER is in remote Protecting administrative state due to a remote FS command, then the SD-P message and the local input are reevaluated as if the LER is in the Normal state. If the LER is in local Protecting administrative state due to a local FS command, then this indication SHALL be ignored (i.e., the indication should have been blocked by the Local Request logic).
- o A remote SD-W message SHALL cause the LER to go into remote Unavailable state and begin transmitting an NR(0,1) message, if the Protecting administrative state is due to a (local or remote) MS command. If the LER is in remote Protecting administrative state due to a remote FS command, then the SD-W message and the local input are reevaluated as if the LER is in the Normal state.

If the LER is in local Protecting administrative state due to a local FS command, then this indication SHALL be ignored

5.11. Updates to Section 4.3.3.4 Protecting Failure State

The bullet item of "Protecting failure state" in Section 3.6. PSC Control States in [RFC6378] includes the degrade condition in Protection failure state. This document follows the same state grouping as [RFC6378] for SD on the working path.

Replace the following text in the first paragraph of Section 4.3.3.4 Protecting Failure State for further clarification on the SD on the working path:

When the protection mechanism has been triggered and the protection domain has performed a protection switch, the domain is in the Protecting failure state. In this state, the normal data traffic SHALL be transported on the protection path. When an LER is in this state, it implies that there either was a local SF condition or it received a remote SF PSC message. The SF condition or message indicated that the failure is on the working path.

This state may be overridden by the Unavailable state triggers, i.e., Lockout of protection or SF on the protection path, or by issuing an FS operator command. This state will be cleared when the SF condition is cleared. In order to prevent flapping due to an intermittent fault, the LER SHOULD employ a Wait-to-Restore timer to delay return to Normal state until the network has stabilized (see Section 3.5).

With:

When the protection mechanism has been triggered and the protection domain has performed a protection switch, the domain is in the Protecting failure state. In this state, the normal data traffic SHALL be transported on the protection path. When an LER is in this state, it implies that there either was a local SF/SD condition or it received a remote SF/SD PSC message. The SF/SD condition or message indicated that the failure/degradation is on the working path.

This state may be overridden by the Unavailable state triggers, i.e., Lockout of protection or SF on the protection path, or by issuing an FS operator command. This state will be cleared when the SF/SD condition is cleared. In order to prevent flapping due to an intermittent fault, the LER SHOULD employ a Wait-to-Restore timer to delay return to Normal state until the network has stabilized (see Section 3.5).

When an LER is in this state due to degradation condition, the user traffic should be duplicated and fed to both working and protection paths if the detection of a SD depends on the presence of user data packets.

Replace the following bullet item text in the transitions in reaction to a local input:

- o A local Clear SF SHALL be ignored if in remote Protecting failure state. If in local Protecting failure state and the LER is configured for revertive behavior, then this input SHALL cause the LER to go into Wait-to-Restore state, start the WTR timer, and begin transmitting a WTR(0,1) message. If in local Protecting failure state and the LER is configured for non-revertive behavior, then this input SHALL cause the LER to go into Do-not-Revert state and begin transmitting a DNR(0,1) message.

With:

- o A local Clear SF for clearing local SF-W SHALL be ignored if in remote Protecting failure state due to remote SF-W. In local Protecting failure state due to local SF-W, clearing local SF-W SHALL cause the LER to go into WTR state, start the WTR timer, and begin transmitting a WTR(0,1) message, if the LER is configured for revertive behavior. Clear local SF-W in local Protecting failure state due to local SF-W SHALL cause the LER to go into Do-not-Revert state and begin transmitting a DNR(0,1) message for non-revertive configuration. In local Protecting Failure state due to local SD-W, if the SF/SD being cleared is SD-W and there is no local SD-P, then go to WTR or DNR state depending on the configuration for revertive behaviour. If there is local SD-P when local SD-W is cleared in local Protecting Failure state due to SD-W, go to local Unavailable state and begin transmitting SD(0.0) message. If the SF/SD being cleared is SD-P in local Protecting Failure due to SD-W, then ignore. In remote Protection Failure state due to remote SD-W, if the SF/SD being cleared is SD-P, then remain in current state and begin transmitting NR(0,1), otherwise, ignore.

Add the following bullet item text to the transitions in reaction to a local input:

- o A local SD-P SHALL be ignored if the LER is in local Protecting Failure state. If in remote Protecting Failure state, the LER SHALL remain in the current state and begin transmission of an SD(0,1) message.
- o A local SD-W SHALL be ignored if the LER is in local Protecting Failure state. If in remote Protecting Failure state, the LER SHALL remain in the current state and begin transmission of an SD(1,1) message.

Add the following SD related sentences to the end of each bullet item text for describing the reaction to remote PSC messages:

remote Lockout of protection: If the LER is in local Protecting Failure state due to local SD-W, then go to remote Unavailable state and begin sending SD(1,0) If in remote Protecting Failure state due to remote SD-W, then go to remote Unavailable state and continue to send the current message with Path=0.

remote Forced Switch: If the LER is in the Protecting Failure state due to local or remote SD-W, go to remote Protecting Administrative state and continue to send the current message.

remote Signal Fail on the protection path: If the LER is in the Protecting Failure state due to local or remote SD-W, go to remote Unavailable state and continue to send the current message with Path=0.

Add the following bullet item text to the transitions in reaction to a remote message:

- o A remote SF-W message received in Protecting Failure state due to local or remote SD-W SHALL cause the LER to remain in Protecting Failure state and continue to send the current message.
- o A remote SD-P message can cause the LER to react differently depending on the cause and locality of current state as follows:
 - * In Protecting Failure state due to remote SF-W, if there is no local request, transition to remote Unavailable state and send NR(0,0). If there is local SD-W input, then transition to remote Unavailable state and send SD(1,0) message. If the local input is SD-P, then transition to local Unavailable state and send SD(0,0) message.

- * In Protecting Failure state due to remote SD-W, if the local input is SD-P, then transition to local Unavailable state. Else, transition to N state.
- * In Protecting Failure state due to local SD-W, if the received SD-P message has Path=1, ignore the message. If the received SD-P message has Path=0 and the active path just before the SD-W is selected as the highest local input was the working path, then go to remote Unavailable state and transmit SD(1,0). If the received SD-P message has Path=0 and the active path just before the SD-W is selected as the highest local input was the protection path, then ignore the received SD-P message.
- o A remote Manual Switch message received in Protecting Failure due to remote SD-W SHALL cause the LER to reevaluate the MS message and local input as if the LER is in the Normal state.

5.12. Updates to Section 4.3.3.5 Wait-to-Restore State

Replace the following paragraph in Section 4.3.3.5 Wait-to-Restore State:

- o When recovering from a failure condition on the working path, the Wait-to-Restore state is used by the PSC protocol to delay reverting to the Normal state, for the period of the WTR timer to allow the recovering failure to stabilize. While in the Wait-to-Restore state, the data traffic SHALL continue to be transported on the protection path. The natural transition from the Wait-to-Restore state to Normal state will occur when the WTR timer expires.

With:

- o When recovering from a failure or degradation condition on the working path, the Wait-to-Restore state is used by the PSC protocol to delay reverting to the Normal state, for the period of the WTR timer to allow the recovering failure/degradation to stabilize. While in the Wait-to-Restore state, the data traffic SHALL continue to be transported on the protection path. The natural transition from the Wait-to-Restore state to Normal state will occur when the WTR timer expires.
- o When an LER is in this state following the recovery of degradation condition, the user traffic will continue to be duplicated and fed to both working and protection paths if the detection of a SD depends on the presence of user data packets.

Add the following bullet item text to the transitions in reaction to a local input:

- o A local SD-P SHALL send the Stop command to the WTR timer, go into local Unavailable state, and begin transmission of an SD(0,0) message.
- o A local SD-W SHALL send the Stop command to the WTR timer, go into local Protecting failure state, and begin transmission of an SD(1,1) message.

Add the following bullet item text to the transitions in reaction to a remote PSC message:

- o A remote SD-P message SHALL send the Stop command to the WTR timer, go into remote Unavailable state, and begin transmission of an NR(0,0) message.
- o A remote SD-W message SHALL send the Stop command to the WTR timer, go into remote Protecting failure state, and begin transmission of an NR(0,1) message.

5.13. Updates to Section 4.3.3.6 Do-not-Revert State

Add the following bullet item text to the transitions in reaction to a local input:

- o A local SD-P SHALL cause the LER to go into local Unavailable state, and begin transmission of an SD(0,0) message.
- o A local SD-W SHALL cause the LER go into local Protecting failure state, and begin transmission of an SD(1,1) message.

Add the following bullet item text to the transitions in reaction to a remote PSC message:

- o A remote SD-P message SHALL cause the LER to go into remote Unavailable state, and begin transmission of an NR(0,0) message.
- o A remote SD-W message SHALL cause the LER to go into remote Protecting failure state, and begin transmission of an NR(0,1) message.

5.14. Updates to Appendix A. PSC State Machine Tables

Add the following extended states:

UA:DP:L Unavailable state due to local SD on protection path
 UA:DP:R Unavailable state due to remote SD-P message
 PF:DW:L Protecting failure state due to local SD on working path
 PF:DW:R Protecting failure state due to remote SD-W message

Add the following default messages:

| State | REQ(FP, P) |
|---------|------------|
| UA:DP:L | SD(0,0) |
| UA:DP:R | NR(0,0) |
| PF:DW:L | SD(1,1) |
| PF:DW:R | NR(0,1) |

Add the following text before the state machine table:

The letter 'r' in the table below stands for reevaluation, and is an indication to reevaluate all inputs (both the local input and the remote message) as if the LER is in the Normal state. See 4.3.3.

Modify the state machine as follows (only modified cells are shown):

Part 1: Local input state machine

| | OC | LO | SF-P | FS | SF-W |
|---------|----|---------|--------|--------|--------|
| N | | | | | |
| UA:LO:L | | | | | |
| UA:P:L | | | | | |
| UA:DP:L | i | UA:LO:L | UA:P:L | PA:F:L | PA:W:L |
| UA:LO:R | | | | | |
| UA:P:R | | | | | |
| UA:DP:R | i | UA:LO:L | UA:P:L | PA:F:L | PF:W:L |
| PF:W:L | | | | | |
| PF:DW:L | i | UA:LO:L | UA:P:L | PA:F:L | PF:W:L |
| PF:W:R | | | | | |
| PF:DW:R | i | UA:LO:L | UA:P:L | PA:F:L | PF:W:L |
| PA:F:L | | | | | |
| PA:M:L | | | | | |
| PA:F:R | | | | | |
| PA:M:R | | | | | |
| WTR | | | | | |

| | | | | | | |
|-----|--|--|--|--|--|--|
| DNR | | | | | | |
|-----|--|--|--|--|--|--|

| | SD-P | SD-W | SF _c | MS | WTRExp |
|---------|---------|---------|-----------------|----|--------|
| N | UA:DP:L | PF:DW:L | | | |
| UA:LO:L | i | i | | | |
| UA:P:L | i | i | [5] | | |
| UA:DP:L | i | i | [20] | i | i |
| UA:LO:R | [21] | [22] | | | |
| UA:P:R | [21] | [22] | | | |
| UA:DP:R | UA:DP:L | [22] | [23] | i | i |
| PF:W:L | i | i | | | |
| PF:DW:L | i | i | [24] | i | i |
| PF:W:R | [25] | [26] | | | |
| PF:DW:R | [25] | PF:DW:L | [27] | i | i |
| PA:F:L | i | i | | | |
| PA:M:L | UA:DP:L | PF:DW:L | | | |
| PA:F:R | [25] | [26] | | | |
| PA:M:R | UA:DP:L | PF:DW:L | | | |
| WTR | UA:DP:L | PF:DW:L | | | |
| DNR | UA:DP:L | PF:DW:L | | | |

Part 2: Remote messages state machine

| | LO | SF-P | FS | SF-W | SD-P | SD-W |
|---------|------|------|------|------|---------|---------|
| N | | | | | UA:DP:R | PF:DW:R |
| UA:LO:L | | | | | i | i |
| UA:P:L | | | | | i | i |
| UA:DP:L | [28] | [29] | [30] | [31] | i | [32] |
| UA:LO:R | | | | | r | r |
| UA:P:R | | | | | r | r |
| UA:DP:R | [33] | [34] | [35] | [36] | i | r |
| PF:W:L | | | | | i | i |
| PF:DW:L | [37] | [38] | [39] | [40] | [41] | i |
| PF:W:R | | | | | [42] | i |
| PF:DW:R | [43] | [44] | [45] | [46] | [47] | i |
| PA:F:L | | | | | i | i |
| PA:M:L | | | | | UA:DP:R | PF:DW:R |
| PA:F:R | | | | | r | r |
| PA:M:R | | | | | UA:DP:R | PF:DW:R |
| WTR | | | | | UA:DP:R | PF:DW:R |

| | | | | | | | |
|--------|---|--------|---|--------|---------|---------|---|
| DNR | | | | | UA:DP:R | PF:DW:R | |
| +----- | + | +----- | + | +----- | + | +----- | + |

| | MS | WTR | DNR | NR |
|---------|----|------|------|----|
| N | | | | |
| UA:LO:L | | | | |
| UA:P:L | | | | |
| UA:DP:L | i | i | i | i |
| UA:LO:R | | | | |
| UA:P:R | | | | |
| UA:DP:R | r | i | i | r |
| PF:W:L | | | | |
| PF:DW:L | i | i | i | i |
| PF:W:R | | | | |
| PF:DW:R | r | [14] | [15] | N |
| PA:F:L | | | | |
| PA:M:L | | | | |
| PA:F:R | | | | |
| PA:M:R | | | | |
| WTR | | | | |
| DNR | | | | |

Replace the following footnote:

- 5 If the SF being cleared is SF-P, transition to N. If it's SF-W, ignore the clear.

With:

- 5 If the SF being cleared is SF-P, transition to N. Otherwise, ignore the clear.

Add the following footnotes for the table:

- 20 If the SF/SD being cleared is SD-P, transition to N. Otherwise, ignore the clear.
- 21 Remain in the current state and transmit SD(0,0).
- 22 Remain in the current state and transmit SD(1,0).

- 23 If the SF/SD being cleared is SD-W, then remain in current state (UA:DP:R) and begin transmitting NR(0,0). Otherwise, ignore the SFC.
- 24 If the SF/SD being cleared is SD-W and there is no local SD-P, then go to WTR or DNR depending on the configuration for revertive behaviour. If there is local SD-P when local SD-W is cleared, go to UA:DP:L state. If the SF/SD being cleared is SD-P then ignore.
- 25 Remain in the current state and transmit SD(0,1).
- 26 Remain in the current state and transmit SD(1,1).
- 27 If the SF/SD being cleared is SD-P, then remain in current state (PF:DW:R) and begin transmitting NR(0,1). Otherwise, ignore.
- 28 Transition to (UA:LO:R) and continue sending SD(0,0)
- 29 Transition to (UA:P:R) and continue sending SD(0,0)
- 30 Transition to (PA:F:R) and send SD(0,1).
- 31 Transition to (PF:W:R) and send SD(0,1)
- 32 If the active path just before the SD is selected as the highest local input was the working path, then ignore. Otherwise, go to PF:DW:R and transmit SD(0,1)
- 33 Transition to (UA:LO:R) state and continue to send the current message.
- 34 Transition to (UA:P:R) state and continue to send the current message.
- 35 Transition to (PA:F:R) state and continue to send the current message with Path=1.
- 36 Transition to (PF:W:R) state and continue to send the current message with Path=1.
- 37 Transition to (UA:LO:R) and send SD(1,0)
- 38 Transition to (UA:P:R) and send SD(1,0)
- 39 Transition to (PA:F:R) and continue to send the current message, SD(1,1).

- 40 Transition to (PF:W:R) and continue to send the current message, SD(1,1).
- 41 If the received SD-P message has Path=1, ignore the message. If the received SD-P message has Path=0 and the active path just before the SD is selected as the highest local input was the working path, then go to UA:DP:R and transmit SD(1,0). If the received SD-P message has Path=0 and the active path just before the SD is selected as the highest local input was the protection path, then ignore the received SD-P message.
- 42 If there is no local request, transition to UA:DP:R and send NR(0,0). If the local input is SD-W, then transition to UA:DP:R and send SD(1,0) message. If the local input is SD-P, then transition to UA:DP:L and send SD(0,0) message.
- 43 Transition to (UA:LO:R) state and continue to send the current message with Path=0.
- 44 Transition to (UA:P:R) state and continue to send the current Message with Path=0.
- 45 Transition to (PA:F:R) state and continue to send the current message.
- 46 Transition to (PF:W:R) state and continue to send the current message.
- 47 If the local input is SD-P, then transition to UA:DP:L. Else, transition to N state.

6. Security considerations

No specific security issue is raised in addition to those ones already documented in [RFC6378]

7. IANA considerations

This document makes no request of IANA.

Note to RFC Editor: this section may be removed on publication as an RFC.

8. Acknowledgements

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.
- [RFC6372] Sprecher, N. and A. Farrel, "MPLS Transport Profile (MPLS-TP) Survivability Framework", RFC 6372, September 2011.
- [RFC6378] Weingarten, Y., Bryant, S., Osborne, E., Sprecher, N., and A. Fulignoli, "MPLS Transport Profile (MPLS-TP) Linear Protection", RFC 6378, October 2011.

9.2. Informative References

- [LIAISON1205] ITU-T SG15, , "Liaison Statement: Recommendation ITU-T G.8131/Y.1382 revision - Linear protection switching for MPLS-TP networks ", <https://datatracker.ietf.org/liaison/1205/> , October 2012.
- [LIAISON1234] ITU-T SG15, , "Liaison Statement: Recommendation ITU-T G.8131 revision - Linear protection switching for MPLS-TP networks ", <https://datatracker.ietf.org/liaison/1234/> , February 2013.

Authors' Addresses

Jeong-dong Ryoo
ETRI
218 Gajeongno
Yuseong-gu, Daejeon 305-700
South Korea

Phone: +82-42-860-5384
Email: ryoo@etri.re.kr

Huub van Helvoort
Huawei Technologies
Karspeldreef 4,
Amsterdam 1101 CJ
the Netherlands

Phone: +31 20 4300832
Email: huub.van.helvoort@huawei.com

Alessandro D'Alessandro
Telecom Italia
via Reiss Romoli, 274
Torino 10141
Italy

Phone: +39 011 2285887
Email: alessandro.dalessandro@telecomitalia.it

Internet Engineering Task Force
Internet-Draft
Intended status: Standards Track
Expires: December 30, 2013

Yimin. Shen
Juniper Networks
Yuji. Kamite
NTT Communications Corporation
Eric. Osborne
Cisco Systems
June 28, 2013

RSVP Setup Protection
draft-shen-mpls-rsvp-setup-protection-03

Abstract

RFC 4090 specifies an RSVP facility-backup fast reroute mechanism for protecting LSPs against link and node failures. This document extends the mechanism to provide so-called "setup protection" for LSPs during their initial Path message signaling time. In particular, it enables a router to reroute an LSP via an existing bypass LSP, when there is a failure of the immediate downstream link or node along the desired path. Therefore, it can be used to avoid LSP signaling failure and reduce setup time in such kind of situation, and allow an LSP to be established temporarily over a bypass LSP when an alternative path can only be resolved at a much later time.

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 December 30, 2013.

Copyright Notice

Copyright (c) 2013 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 | 2 |
| 2. Specification of Requirements | 4 |
| 3. Theory of Operation | 4 |
| 3.1. New RSVP Attribute Flag | 5 |
| 3.2. New RSVP Attributes TLVs | 5 |
| 3.2.1. Protected LSP Sender IPv4 Address TLV | 6 |
| 3.2.2. Protected LSP Sender IPv6 Address TLV | 6 |
| 3.3. PLR behavior | 7 |
| 3.4. MP behavior | 9 |
| 3.5. Local Revertive Mode | 9 |
| 4. IANA Considerations | 10 |
| 5. Security Considerations | 10 |
| 6. Acknowledgements | 10 |
| 7. References | 10 |
| 7.1. Normative References | 10 |
| 7.2. Informative References | 11 |
| Authors' Addresses | 11 |

1. Introduction

In RSVP facility-backup fast reroute (FRR) [RFC 4090], the router at a point of local repair (PLR) of an LSP can redirect traffic via a bypass LSP upon a failure of the immediate downstream link or node. Such protection is normally established after the LSP has been set up. This is because the PLR must know the label and address of the next-hop router (in link protection) or those of the next-next-hop router (in node protection), before it can select or create a bypass LSP to protect the LSP. The information of the label and the address is carried in the Resv message of the LSP.

Imagine a scenario where a new LSP is being signaled, and its Path message carries an EXPLICIT_ROUTE object (ERO) with a strict path that is statically configured or computed offline based on a topology that assumes no failure of the network. If a link or node along the path happens to be in a failure condition, RSVP signaling will stop at the router upstream adjacent to the failure, as the next hop in the strict path no longer matches the current network topology. This will be the case even if there is an existing bypass LSP protecting the link or node for some existing LSPs. In other words, this new LSP is not protected during the setup time, i.e. the initial Path message signaling.

In this situation, the network would normally rely on IGP to update traffic engineering (TE) information throughout the network, and the router upstream adjacent to the failure to send a PathErr message to trigger the ingress router to compute and signal a new path. However, this approach may not always be possible or desirable in the following scenarios:

1. Static strict path. As described above, if the ERO carries an explicit path with a sequence of strict hops that are statically configured or computed offline based on a topology assuming no network failure, the LSP will not be established until the path is modified. This is a typical case where CSPF calculation is disabled at the LSP's ingress router due to the operational preference of service provider.
2. LSPs with a strict requirement for setup time. IGP TE information flooding, PathErr message propagation and path re-computation and re-signaling may introduce a significant delay to LSP establishment. This may impact on LSP setup time, and even become unacceptable for LSPs that have a strict requirement for it, such as on-demand transport LSPs for real-time data or TV broadcast. For these LSPs, a guaranteed establishment and setup time are considered as more important than path optimality.
3. Sibling P2MP sub-LSPs sharing a common link. In this case, the new LSP is a sub-LSP of a P2MP LSP, and its desired path is supposed to share the failed link with an existing sibling sub-LSP, i.e. another sub-LSP of the same P2MP LSP, which is being protected by a bypass LSP. If the new sub-LSP is rerouted via a different path, it will not be able to share the data flow over the bypass LSP with that sibling sub-LSP, creating unnecessary traffic flow in the network.

For networks where a failure, delay or resignaling during LSP setup is not desirable, this document extends the RSVP facility-backup fast reroute mechanism to provide a graceful solution, called "setup

protection". During the initial Path message signaling of an LSP, if there is a link or node failure along the desired path, and if there is a bypass LSP protecting the link or node, the LSP can be signaled through the bypass LSP without a delay. The LSP will be established as if it were originally set up along the desired path (i.e. primary path) and then failed over to the bypass LSP after the failure. Meanwhile, actions may be taken to resolve the failure or resignal the LSP via an alternative path, by following procedures or timing appropriate to the service provider. The setup protection is applicable to both P2P LSPs and P2MP LSPs, when such kind of temporary rerouting is not considered as a violation of desired path, as in the case of the normal fast reroute. It may be enabled by policy on a per LSP basis.

2. Specification of Requirements

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119.

3. Theory of Operation

When an LSP is being signaled by RSVP, a Path message is sent hop by hop from the ingress router to the egress router, following the path defined by an ERO. The setup protection mechanism in this document enables a router to reroute the LSP via a bypass LSP, if the router detects a failure of the immediate downstream link or node represented by the next hop in the ERO, called "next ERO hop". In this case, the current router is referred to as a PLR.

The mechanism is only relevant when the Path message carries the "local protection desired" flag in the SESSION_ATTRIBUTE object [RFC 4090] and a new "setup protection desired" flag defined in this document (Section 3.1). That is, setup protection is explicitly requested for the LSP.

In link protection, the mechanism is only applicable when the next ERO hop received by a PLR is a strict hop. In node protection, the mechanism is only applicable when both the next and the next-next ERO hops received by the PLR are strict hops. Otherwise, setup protection would be unnecessary, as the router may perform a loose hop expansion to reroute the LSP via any alternative path around the downstream failure. The strict ERO hops ensure that the PLR can unambiguously decide the intended downstream link or node and reliably detect its status. In link protection, the strict next ERO hop also indicates the merge point (MP), i.e. the destination of the bypass LSP to be used to reroute the LSP. In node protection, the strict next-next ERO hop indicates the MP.

When performing setup protection, the PLR signals a backup LSP by tunneling Path message through the bypass LSP. Like the Path message of a backup LSP in the normal facility-backup FRR ([RFC 4090]), this Path message carries an address of the PLR as the sender address in SENDER_TEMPLATE object. In addition, the Path message also carries the information of the protected LSP (Section 3.2). When the MP receives the Path message, it terminates the backup LSP, and re-creates the protected LSP. If the MP is the egress router of the protected LSP, it terminates the protected LSP as well. If the MP is a transit router of the protected LSP, it signals the LSP further downstream.

Eventually, the LSP will be established end to end, with the backup LSP tunneled through the bypass LSP from the PLR to the MP. The RSVP state on the PLR and the MP and the RSVP messages generated by these routers are no different than those in a post-failure situation of a normal facility-backup FRR.

Later, when the failure is resolved, the PLR MAY revert the LSP to the primary path, in the same manner as the local revertive mode specified in [RFC 4090].

The setup protection MAY be enabled and disabled on a router based on configuration. For an LSP to be setup-protected, the mode MUST be enabled on both PLR and MP. If it is enabled on the PLR but disabled on the MP, the MP SHOULD reject the Path message of the backup LSP and send a PathErr message, as described Section 3.4.

3.1. New RSVP Attribute Flag

In order for an LSP to explicitly request setup protection, this document defines a new "setup protection desired" flag for the Attribute Flags TLV of the LSP_ATTRIBUTES object [RFC5420]. The flag is set by the ingress router in the Path message of the LSP, i.e. the protected LSP. It MUST be supported by all routers that intend to serve as PLRs for setup protection.

3.2. New RSVP Attributes TLVs

This document defines the following two new RSVP Attributes TLVs [RFC 5420]. They are used by a PLR to convey to an MP the original sender address in SENDER_TEMPLATE object of the Path message of a protected LSP.

- o Protected LSP Sender IPv4 Address TLV
- o Protected LSP Sender IPv6 Address TLV

One of the TLVs SHOULD be carried by the LSP_REQUIRED_ATTRIBUTES object of the Path message of the backup LSP that the PLR sends to the MP. The information is used by the MP to build Path message for the protected LSP. The MP SHOULD NOT propagate the TLV downstream via that Path message.

3.2.1. Protected LSP Sender IPv4 Address TLV

The Protected LSP Sender IPv4 Address TLV is defined with type TBD. It is allowed in LSP_REQUIRED_ATTRIBUTES object, and not allowed in LSP_ATTRIBUTES object. The encoding is as below.

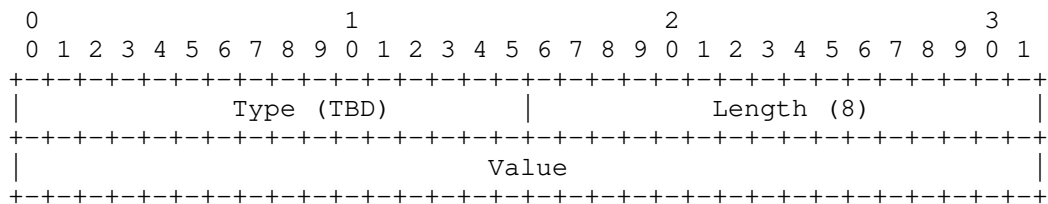


Figure 1

Type

TBD

Length

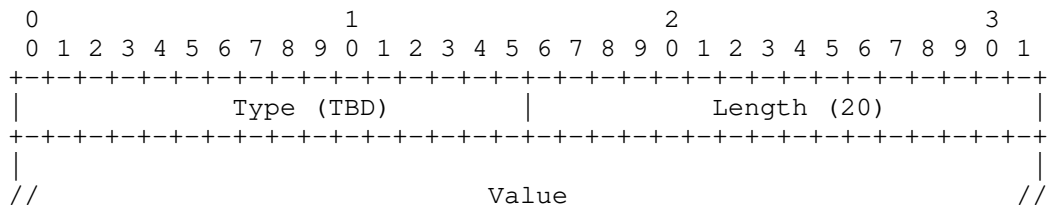
8

Value

Original sender address in the IPv4 SENDER_TEMPLATE object of the protected LSP.

3.2.2. Protected LSP Sender IPv6 Address TLV

The Protected LSP Sender IPv6 Address TLV is defined with type TBD. It is allowed in LSP_REQUIRED_ATTRIBUTES object, and not allowed in LSP_ATTRIBUTES object. The encoding is as below.



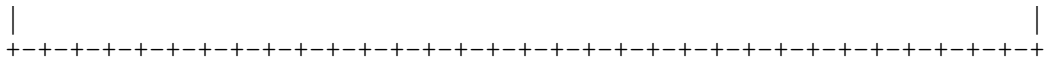


Figure 2

Type

TBD

Length

20

Value

Original sender address in the IPv6 SENDER_TEMPLATE object of the protected LSP.

3.3. PLR behavior

When a router has a Path message to send out, if the Path message carries the "local protection desired" flag in the SESSION_ATTRIBUTE object and the "setup protection desired" flag in the LSP_ATTRIBUTES object, and if the next ERO hop is a strict IPv4 or IPv6 prefix, the router SHOULD validate the reachability of the prefix against routing tables, traffic engineering (TE) database, or a database that reflects the current status of the network topology. If the prefix is reachable and is one hop away from the current router, the router should send out the Path message as it is. Otherwise, there is a possibility that the link or node corresponding to the prefix has failed.

The router SHOULD further search for an existing bypass LSP that is protecting the prefix. If the protected LSP desires link protection, the destination of the bypass LSP (i.e. MP) must be the router that owns the prefix. If the LSP desires node protection and the next-next ERO hop of the LSP is a strict prefix, the MP must be the router that owns this prefix.

If a bypass LSP is not found by the above criteria, the router MUST originate a PathErr with code = 24 (routing problem) and sub-code = 2 (bad strict node).

If a bypass LSP is found, the router **MUST** act as a PLR for setup protection, and reroute the protected LSP via the bypass LSP. If multiple satisfactory bypass LSPs exist, the PLR **MAY** select one based on bandwidth constraints or local policies. Specifically, if the protected LSP is a sub-LSP of a P2MP LSP, a bypass LSP that is

protecting an existing sibling sub-LSP MUST be preferred, to minimize traffic duplication in the network.

The PLR SHOULD NOT send the Path message of the protected LSP any further. Instead, it MUST create a backup LSP, and send a Path message of the backup LSP to the MP via the bypass LSP. The Path message is constructed by using the sender template specific method [RFC 4090]. In particular, it has the sender address in the SENDER_TEMPLATE object set to an address of the PLR. It MUST carry an LSP_REQUIRED_ATTRIBUTES object with a Protected LSP Sender IPv4 Address TLV or Protected LSP Sender IPv6 Address TLV.

Upon receiving a Resv message of the backup LSP from the MP, the PLR SHOULD bring up both of the backup LSP and the protected LSP. If the PLR is the ingress router of the protected LSP, the LSP has been set up successfully. If the PLR is a transit router, it MUST send a Resv message upstream for the protected LSP, with the "local protection available" and "local protection in use" set to 1, and if applicable, the "node protection" and "bandwidth protection" flags set to 1, in the RRO hop corresponding to the PLR. The PLR SHOULD also originate a PathErr message with code = 25 (notify error) and sub-code = 3 (tunnel locally repaired), as if the LSP had just fell over to the bypass LSP.

The PLR SHOULD also install a forwarding entry for the protected LSP. In the typical case, the forwarding entry should result in two outgoing labels for packets. The inner label is the backup LSP's label, and the outer label is the bypass LSP's label. However, the forwarding entry may result in one or no label, if either or both of the backup LSP and the bypass LSP have the Implicit NULL label.

If the PLR receives a PathErr message when signaling the backup LSP, the PLR MUST NOT bring up the backup LSP or the protected LSP. If the PLR is a transit router of the protected LSP, it MUST propagate the PathErr message upstream for the protected LSP. Likewise, if the PLR receives a PathErr message of the backup LSP after the backup LSP and the primary LSP have previously been brought up, and the PLR is a transit router of the protected LSP, it SHOULD also propagate the PathErr message upstream for the protected LSP.

When the PLR receives a ResvTear message of the backup LSP, the PLR MUST bring down both the backup LSP and the protected LSP. If the PLR is a transit router of the protected LSP, it MUST send a ResvTear message upstream for the protected LSP.

In any cases where the PLR needs to bring down the protected LSP due to a received PathTear message, an RSVP state time-out, a configuration change, an administrative command, etc, the PLR MUST

also bring down the backup LSP by sending a PathTear message through the bypass LSP.

3.4. MP behavior

When an MP receives the Path message of a backup LSP, it MUST realize the setup protection situation based on the presence of Protected LSP Sender IPv4 Address TLV or Protected LSP Sender IPv6 Address TLV in LSP_REQUIRED_ATTRIBUTES object.

If setup protection mode is disabled on the MP, it MUST reject the Path message, by sending a PathErr with code = 2 (policy control failure) to the PLR.

Otherwise, the MP MUST terminate the backup LSP and re-create the protected LSP. If the MP is the egress router of the protected LSP, it MUST also terminate the protected LSP. If the MP is a transit router of the LSP, it MUST send a Path message downstream for the protected LSP. The Path message has the sender address in SENDER_TEMPLATE object set to the original address of the ingress router, based on the above received Protected LSP Sender IPv4 Address TLV or Protected LSP Sender IPv6 Address TLV. The Path message MUST NOT carry any Protected LSP Sender IPv4 Address TLV or Protected LSP Sender IPv6 Address TLV in LSP_REQUIRED_ATTRIBUTES object.

The MP MUST allocate a label for the backup LSP, and distribute it to the PLR via Resv message of the backup LSP. If the protected LSP is a sub-LSP of a P2MP LSP and there is an existing sibling sub-LSP whose backup LSP is tunneled through the same bypass LSP, the MP MUST allocate the same label as the sibling sub-LSP, in order to avoid traffic duplication at the PLR.

When the MP receives a PathTear message for the backup LSP, it MUST bring down both the backup LSP and the protected LSP. If the MP is a transit router of the protected LSP, it MUST send a PathTear message downstream for the protected LSP.

In any cases where the MP receives or originates a PathErr or ResvTear message for the protected LSP, the MP MUST send the same type of message to the PLR for the backup LSP.

3.5. Local Revertive Mode

When the failed link or node is restored, the PLR MAY revert the protected LSP to its desired primary path, by following the procedure of local revertive mode described in [RFC 4090].

4. IANA Considerations

This document defines a new flag for the Attribute Flags TLV, which is carried in the LSP_ATTRIBUTES Object of Path message. This flag is used to communicate whether setup protection is desired for an LSP. The value of the new flag needs to be assigned by IANA.

Setup Protection Desired: TBD

This document defines two new RSVP Attributes TLVs, which are carried in the LSP_REQUIRED_ATTRIBUTES object of Path message. The values of the new types need to be assigned by IANA.

Protected LSP Sender IPv4 Address TLV

Protected LSP Sender IPv6 Address TLV

5. Security Considerations

The security considerations discussed in RFC 3209, RFC 4090 and RFC 4875 apply to this document.

6. Acknowledgements

Thanks to Rahul Aggarwal, Disha Chopra, and Nischal Sheth for their contribution.

7. References

7.1. Normative References

- [RFC2205] Braden, B., Zhang, L., Berson, S., Herzog, S., and S. Jamin, "Resource ReSerVation Protocol (RSVP) -- Version 1 Functional Specification", RFC 2205, September 1997.
- [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.
- [RFC4090] Pan, P., Swallow, G., and A. Atlas, "Fast Reroute Extensions to RSVP-TE for LSP Tunnels", RFC 4090, May 2005.
- [RFC5420] Farrel, A., Papadimitriou, D., Vasseur, JP., and A. Ayyangarps, "Encoding of Attributes for MPLS LSP Establishment Using Resource Reservation Protocol Traffic Engineering (RSVP-TE)", RFC 5420, February 2009.

- [RFC4875] Aggarwal, R., Papadimitriou, D., and S. Yasukawa, "Extensions to Resource Reservation Protocol - Traffic Engineering (RSVP-TE) for Point-to-Multipoint TE Label Switched Paths (LSPs)", RFC 4875, May 2007.
- [RFC3471] Berger, L., "Generalized Multi-Protocol Label Switching (GMPLS) Signaling Functional Description", RFC 3471, January 2003.
- [RFC3472] Ashwood-Smith, P. and L. Berger, "Generalized Multi-Protocol Label Switching (GMPLS) Signaling Constraint-based Routed Label Distribution Protocol (CR-LDP) Extensions", RFC 3472, January 2003.
- [RFC3031] Rosen, E., Viswanathan, A., and R. Callon, "Multiprotocol Label Switching Architecture", RFC 3031, January 2001.

7.2. Informative References

- [RFC5920] Fang, L., "Security Framework for MPLS and GMPLS Networks", RFC 5920, July 2010.

Authors' Addresses

Yimin Shen
Juniper Networks
10 Technology Park Drive
Westford, MA 01886
USA

Phone: +1 9785890722
Email: yshen@juniper.net

Yuji Kamite
NTT Communications Corporation
Granpark Tower 3-4-1 Shibaura, Minato-ku
Tokyo 108-8118
Japan

Email: y.kamite@ntt.com

Eric Osborne
Cisco Systems

Email: eosborne@cisco.com

MPLS Working Group
Internet-Draft
Intended status: Standards Track
Expires: January 9, 2014

IJ. Wijnands, Ed.
Cisco
A. Gulko
Thomson Reuters
U. Joorde
Deutsche Telekom
J. Tantsura
Ericsson
July 8, 2013

mLDP in-band signalling Wildcard encoding
draft-wijnands-mpls-mldp-in-band-wildcard-encoding-00

Abstract

Documents [RFC6826] and [I-D.ietf-l3vpn-mldp-vrf-in-band-signaling] define a solution to splice an IP multicast tree together with a multipoint LSP in the global or VRF context. In these drafts the Multipoint Label Distribution Protocol (mLDP) Opaque TLV encodings have been documented for Source specific and Bidir IP multicast trees. For each IP multicast tree a multipoint LSP is created. There are scenarios where it is beneficial to support shared trees and allow aggregation such that fewer multipoint LSPs are created in the network. This document defines wildcard encodings to be used for the Source or Group fields of the existing opaque encodings. With the wildcard encoding it is possible to create a single multipoint LSP that is used to represent *all* sources for a given multicast group or *all* groups for a given source.

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 January 9, 2014.

Copyright Notice

Copyright (c) 2013 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. Terminology and Definitions | 4 |
| 2. Introduction | 4 |
| 3. PIM shared tree forwarding | 5 |
| 4. IGMP/MLD proxying | 6 |
| 5. Selective Source mapping | 6 |
| 6. Wildcard Source | 6 |
| 7. Wildcard Group | 7 |
| 8. Root node address discovery | 8 |
| 9. Anycast RP | 8 |
| 10. Wildcard encoding | 8 |
| 11. Acknowledgements | 9 |
| 12. IANA Considerations | 9 |
| 13. Security Considerations | 9 |
| 14. References | 9 |
| 14.1. Normative References | 9 |
| 14.2. Informative References | 9 |
| Authors' Addresses | 10 |

1. Terminology and Definitions

PIM: Protocol Independent Multicast.

IGMP: Internet Group Management Protocol.

MLD: Multicast Listener Discovery.

IP multicast tree: An IP multicast distribution tree identified by a IP multicast group address and optionally a Source IP address, also referred to as (S,G) and (*,G).

MP-LSP: A P2MP or MP2MP LSP.

PIM-ASM: PIM Any Source Multicast.

PIM-SSM: PIM Source Specific Multicast.

PIM-SM: PIM Sparse-mode Multicast.

RP: The PIM Rendezvous Point.

mLDP: Multipoint LDP.

In-band signaling: Using the opaque value of a mLDP FEC element to carry the (S,G) or (*,G) identifying a particular IP multicast tree.

Ingress LSR: Source of the P2MP LSP, also referred to as root node.

Egress LSR: A LSR that has receivers attached, also referred to as leaf node.

Threshold Infinity: A PIM-SM procedure where no source specific multicast (S,G) trees are created for multicast packets that are forwarded down the shared tree (*,G).

2. Introduction

Documents [RFC6826] and [I-D.ietf-l3vpn-mlbp-vrf-in-band-signaling] define a solution to splice an IP multicast tree together with a multipoint LSP in the global or VRF context. In these drafts the Multipoint Label Distribution Protocol (mLDP) Opaque TLV encodings have been documented for Source specific and Bidir IP multicast trees. For each IP multicast tree a mLDP MP-LSP is created. There are scenarios where it is beneficial to support shared trees and allow aggregation such that fewer multipoint LSPs are created in the

network. This document defines wildcard encodings that can be used in Source or Group fields of the existing Opaque TLV encodings. With the wildcard encoding it is possible to create a single multipoint LSP used to represent **all** sources for a given multicast group or **all** groups for a given source.

The behaviour of an mLDP in-band signalled multipoint LSPs containing a wildcard entry follows the procedures defined in [RFC6826] and [I-D.ietf-l3vpn-mlbp-vrf-in-band-signaling]. This draft does not talk about already defined procedures but only documents the differences.

There are a few scenarios (not limited to) where wildcard encoding is useful, for example;

- o PIM Shared tree forwarding with threshold infinity.
- o IGMP/MLD proxying.
- o Selective Source mapping.

These scenarios are discussed in this draft below.

3. PIM shared tree forwarding

PIM [RFC4601] has the concept of a shared tree, known as (*,G). This means, **all** Sources for a given Group in the ASM range. The (*,G) is built towards the Rendezvous Point (RP) that typically joins **all** multicast sources for this group. The RP will then forward all the IP multicast packets for this group down the (*,G) tree towards the receivers. There are several procedures how the RP learns about these sources, for example; PIM Registers [RFC4601], MSDP [RFC3618] or a Source that is directly connected to the RP. In some cases, the last hop routers does not wish to join the source trees, and expect to receive all the traffic for group G from the (*,G) tree; in this case, we say that the last hop routers have 'threshold infinity' for group G. This is optional behaviour documented in the [RFC4601]. This is often used in deployments where the RP is between the multicast sources and the multicast receivers for group G, i.e., the shortest path from any source to any receiver of the group goes through the RP. In this scenario, there is no advantage for a last hop router to join a source tree for the group, since joining a source tree would not change the path of the multicast data from the source. The only effect of executing the complicated procedures for joining a source tree and pruning the source off the shared tree would be to an increase of the amount of multicast routing state.

This deployment model can be implemented using wildcards. The egress router will splice the (*,G) IP multicast tree to a mLDP Multipoint LSP where the Source address is encoded as wildcard entry. In scenarios where it does not make sense to apply "threshold infinity" to a given ASM group, a more complex set of procedures are needed, as per [I-D.rekhter-pim-sm-over-mldp].

4. IGMP/MLD proxying

There are scenarios where the multicast senders and receivers are directly connected to a MPLS routing domain and mLDP is available. In these cases we can apply "IGMP/MLD proxying" and avoid using PIM as a multicast routing protocol to transport multicast packets from the senders to the receivers. The senders and receivers consider the MPLS domain to be single hop between each other. [RFC4605] documents procedures where a multicast routing protocol is not necessary to build a 'simple tree'. Within the MPLS domain mLDP will be used to build a 'spanning tree' to avoid looping and duplication of packets, but for the point of view of the senders and receivers this is hidden. The procedures as defined [RFC4605] are applicable since the senders and receivers are considered to be one hop away from each other.

For mLDP to build a tree, it needs to know the root of the tree. Following the procedures as defined in [RFC4605] we depend on manual configuration of the mLDP root for the ASM multicast group. The Source will be encoded as a wildcard entry.

5. Selective Source mapping

In IPTV deployments, rather often, the content servers are co-located in a few sites. Popular channels are often statically configured and always forwarded over the core MPLS network to the egress routers. Since these channels are statically defined, they MAY also be forwarded over a multipoint LSP with wildcard encoding. The sort of wildcard encoding that needs to be used (Source and/or Group) depends on the Source/Group allocation policy of the IPTV provider. Other options are to use MSDP [RFC3618] or BGP AD [RFC6513] for source discovery by the ingress LSR. Based on the received wildcard, the ingress LSR can make a selection out of the IP multicast streams it has state for.

6. Wildcard Source

When the IP multicast component on the ingress LSR has received a

wildcard source from mLDP it may have been initiated by one of the scenarios described in this draft. How the wildcard source is to be interpreted is a local matter and follows the rules below;

1. If PIM is enabled and the group is a non-bidirectional ASM group, the wildcard source is treated as having received a (*,G) IGMP/MLD report from a downstream node and the procedures as defined in [RFC4601] are followed.
2. If PIM is enabled and the group mode is PIM-SSM, all multicast sources known for the group on the root node must be forwarded down the multipoint LSP.
3. If PIM is not enabled for this group, the wildcard source is treated as having received a (*,G) IGMP/MLD report from a downstream node and the procedures as defined in [RFC4605] are followed.

The IP multicast component on the egress LSR determines when a Wildcard Source is to be used in a mLDP Opaque TLV encoding. How the IP multicast component determines this is a local matter and potentially subjected to explicit user configuration. It MAY however use the following rules (with or without explicit user configuration);

1. If PIM is enabled, the group is a non-bidirectional ASM group and the RP is reachable via a BGP route, a Wildcard Source encoding MAY be used to signal group membership (*,G) to the ingress LSR, using the BGP next hop as the ingress LSR (root of the LSP). Also see Section 8.
2. If PIM is not enabled for this group and an IGMP/MLD group membership report has been received, the IP multicast component may use a Wildcard Source encoding to signal the group membership (*,G) to a Proxy device (root of the LSP). The procedures how to determine the Proxy device for a given group are defined in [RFC4605].

The wildcard source encoding MUST NOT appear in the "Bidir TLVs" that are defined in [RFC6826] sections 3.3 and 3.4."

A wildcard group in combination with a wildcard source encoding is under investigation.

7. Wildcard Group

When the IP multicast component on the root node receives a wildcard

group encoding, the root node SHOULD apply the wildcard encoding to the existing IP multicast routing table and forward all the IP multicast stream(s) that match the given Source. Note, this behaviour is independent of the PIM group mode (ie. ASM or SSM).

The IP multicast component on the egress LSR determines when a Wildcard Group is to be used in a mLDP Opaque TLV encoding. How the IP multicast component determines this is a local matter and subjected to explicit user configuration.

The wildcard group encoding for PIM bidir is under investigation.

A wildcard source in combination with a wildcard group encoding is under investigation.

8. Root node address discovery

Documents [RFC6826] and [I-D.ietf-l3vpn-mlbp-vrf-in-band-signaling] describe procedures to discover the mLDP root node address by using the Source of the IP multicast stream. When a wildcard source encoding is used, PIM is enabled and the group is a non-bidirectional ASM group, a similar procedure is applied. The only difference with the above mentioned procedures is that the Proxy device or RP address is used instead of the Source to discover the mLDP root node address.

In all other cases some sort of manual configuration is applied in order to find the root node. Note, finding the root node is a local implementation matter and not limited to the solutions mentioned in this draft.

9. Anycast RP

With in-band signalling there is likely no RP to Group mappings distribution taking place over the MPLS core to the different IP multicast sites. The RP address is likely statically configured on each multicast site. In these cases it makes sense to configure an Anycast RP Address to provide redundancy. See [RFC3446] for more details.

10. Wildcard encoding

The source and group fields in the Transit IPv4, IPv6, VPNv4 and VPNv6 Source TLVs, as documented in [RFC6826] and [I-D.ietf-l3vpn-mlbp-vrf-in-band-signaling] only allow valid IP addresses to be encoded. This document proposes to use a source/

group field of *all* zero's to be used as wildcard encoding.

11. Acknowledgements

The authors would like to thank Eric Rosen for his valuable comments.

12. IANA Considerations

There are no new allocations required from IANA.

13. Security Considerations

There are no security considerations other than ones already mentioned in [RFC6826] and [I-D.ietf-l3vpn-mldp-vrf-in-band-signaling].

14. References

14.1. Normative References

- [I-D.ietf-l3vpn-mldp-vrf-in-band-signaling]
Wijnands, I., Hitchen, P., Leymann, N., Henderickx, W.,
and a. arkadiy.gulko@thomsonreuters.com, "Multipoint Label
Distribution Protocol In-Band Signaling in a VRF Context",
draft-ietf-l3vpn-mldp-vrf-in-band-signaling-01 (work in
progress), June 2013.
- [RFC4605] Fenner, B., He, H., Haberman, B., and H. Sandick,
"Internet Group Management Protocol (IGMP) / Multicast
Listener Discovery (MLD)-Based Multicast Forwarding
("IGMP/MLD Proxying")", RFC 4605, August 2006.
- [RFC6826] Wijnands, IJ., Eckert, T., Leymann, N., and M. Napierala,
"Multipoint LDP In-Band Signaling for Point-to-Multipoint
and Multipoint-to-Multipoint Label Switched Paths",
RFC 6826, January 2013.

14.2. Informative References

- [I-D.rekhter-pim-sm-over-mldp]
Rekhter, Y. and R. Aggarwal, "Carrying PIM-SM in ASM mode
Trees over P2MP mLDP LSPs",
draft-rekhter-pim-sm-over-mldp-04 (work in progress),
August 2011.

- [RFC3446] Kim, D., Meyer, D., Kilmer, H., and D. Farinacci, "Anycast Rendezvous Point (RP) mechanism using Protocol Independent Multicast (PIM) and Multicast Source Discovery Protocol (MSDP)", RFC 3446, January 2003.
- [RFC3618] Fenner, B. and D. Meyer, "Multicast Source Discovery Protocol (MSDP)", RFC 3618, October 2003.
- [RFC4601] Fenner, B., Handley, M., Holbrook, H., and I. Kouvelas, "Protocol Independent Multicast - Sparse Mode (PIM-SM): Protocol Specification (Revised)", RFC 4601, August 2006.
- [RFC6513] Rosen, E. and R. Aggarwal, "Multicast in MPLS/BGP IP VPNs", RFC 6513, February 2012.

Authors' Addresses

IJsbrand Wijnands (editor)
Cisco
De kleetlaan 6a
Diegem, 1831
Belgium

Phone:
Email: ice@cisco.com

Arkadiy Gulko
Thomson Reuters
195 Broadway
New York NY 10007
USA

Email: arkadiy.gulko@thomsonreuters.com

Uwe Joorde
Deutsche Telekom
Hammer Str. 216-226
Muenster D-48153
DE

Email: Uwe.Joorde@telekom.de

Jeff Tantsura
Ericsson
300 Holger Way
San Jose, California 95134
USA

Email: Jeff.Tantsura@ericsson.com

MPLS Working Group
Internet-Draft
Intended status: Standards Track
Expires: November 07, 2013

H. van Helvoort, Ed.
Huawei Technologies
J. Ryoo, Ed.
ETRI
H. Zhang
Huawei Technologies
F. Huang
Alcatel-Lucent Shanghai Bell
H. Li
China Mobile
A. D'Alessandro
Telecom Italia
May 06, 2013

Linear Protection Switching in MPLS-TP
draft-zulr-mpls-tp-linear-protection-switching-07.txt

Abstract

This document specifies a linear protection switching mechanism for MPLS-TP. This mechanism supports 1+1 unidirectional/bidirectional protection switching and 1:1 bidirectional protection switching. It is purely supported by MPLS-TP data plane, and can work without any control plane.

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 as defined by the ITU-T.

Status of This Memo

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

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

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

This Internet-Draft will expire on November 07, 2013.

Copyright Notice

Copyright (c) 2013 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. Linear protection switching overview | 4 |
| 2.1. Protection architecture types | 4 |
| 2.1.1. 1+1 architecture | 4 |
| 2.1.2. 1:1 architecture | 4 |
| 2.1.3. 1:n architecture | 4 |
| 2.2. Protection switching type | 5 |
| 2.3. Protection operation type | 5 |
| 3. Protection switching trigger conditions | 5 |
| 3.1. Fault conditions | 6 |
| 3.2. External commands | 6 |
| 3.2.1. End-to-end commands | 6 |
| 3.2.2. Local commands | 7 |
| 4. Protection switching schemes | 7 |
| 4.1. 1+1 unidirectional protection switching | 7 |
| 4.2. 1+1 bidirectional protection switching | 8 |
| 4.3. 1:1 bidirectional protection switching | 9 |
| 5. APS protocol | 10 |
| 5.1. APS PDU format | 10 |
| 5.2. APS transmission | 13 |
| 5.3. Hold-off timer | 13 |
| 6. Protection switching logic | 14 |
| 7. Protection switching state transition table | 16 |
| 8. Security considerations | 18 |
| 9. IANA considerations | 18 |
| 10. Acknowledgements | 18 |
| 11. References | 18 |
| 11.1. Normative References | 18 |
| 11.2. Informative References | 18 |

| | |
|--|----|
| Appendix A. Operation examples of APS protocol | 18 |
| Authors' Addresses | 24 |

1. Introduction

MPLS-TP is defined as transport profile of MPLS technology to fulfill the deployment in transport network. A typical feature of transport network is that it can provide fast protection switching for end-to-end or segments. The protection switching time is generally required to be less than 50ms according to the strictest requirement of services such as voice, private line, etc.

The goal of linear protection switching mechanism is to satisfy the requirement of fast protection switching for MPLS-TP network. Linear protection switching means that, for one or more working transport entities, there is one protection transport entity, which is disjoint from any of working transport entities, ready for taking over the service transmission when a working transport entity failed.

This document specifies 1+1 unidirectional protection switching mechanism for unidirectional transport entity (either point-to-point or point-to-multipoint) as well as bidirectional point-to-point transport entity, and 1+1/1:1 bidirectional protection switching mechanism for point-to-point bidirectional transport entity. Since bidirectional protection switching needs the coordination of the two endpoints of the transport entity, this document also specifies APS (Automatic Protection Switching) protocol details which is used for this purpose.

The linear protection mechanism described in this document is applicable to both LSPs and PWs.

The APS protocol specified in this document is based on the same principles and behavior of the APS protocol designed for SONET/SDH networks (i.e., it is mature and proven) and provides commonality with the established operation models utilized in other transport network technologies (e.g., SDH/SONET and OTN).

It is also worth noting that multi-vendor implementations of the APS protocol described in this document already exist.

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 as defined by the ITU-T.

2. Linear protection switching overview

To guarantee the protection switching time, for a working transport entity, its protection transport entity is always pre-configured before the failure occurs. Normally, the normal traffic will be transmitted and received on the working transport entity. The switching to protection transport entity is usually triggered by link /node failure, external commands, etc. Note that external commands are often used in transport network by operators, and they are very useful in cases of service adjustment, path maintenance, etc.

2.1. Protection architecture types

2.1.1. 1+1 architecture

In the 1+1 architecture, a protection transport entity is associated with the working transport entity. The normal traffic is permanently bridged onto both the working transport entity and the protection transport entity at the source endpoint of the protected domain. The normal traffic on working and protection transport entities is transmitted simultaneously to the sink endpoint of the protected domain where a selection between the working and protection transport entity is made, based on predetermined criteria, such as signal fail and signal degrade indications.

2.1.2. 1:1 architecture

In the 1:1 architecture, a protection transport entity is associated with the working transport entity. When the working transport entity is determined to be impaired, the normal traffic must be transferred from the working to the protection transport entity at both the source and sink endpoints of the protected domain. The selection between the working and protection transport entities is made based on predetermined criteria, such as signal fail and signal degrade indications from the working or protection transport entity.

The bridge at source endpoint can be realized in two ways: it is either a selector bridge or a broadcast bridge. With a selector bridge the normal traffic is connected either to the working transport entity or the protection transport entity. With a broadcast bridge the normal traffic is permanently connected to the working transport entity, and in case a protection switch is active also to the protection transport entity. Broadcast bridge is recommended to be used in revertive mode only.

2.1.3. 1:n architecture

Details for the 1:n protection switching architecture will be provided in a future version of this draft.

It is worth noting that the APS protocol defined here is ready to support 1:n operations.

2.2. Protection switching type

The linear protection switching types can be a unidirectional switching type or a bidirectional switching type.

- o Unidirectional switching type: Only the affected direction of working transport entity is switched to protection transport entity; the selectors at each endpoint operate independently. This switching type is recommended to be used for 1+1 protection in this document.
- o Bidirectional switching type: Both directions of working transport entity, including the affected direction and the unaffected direction, are switched to protection transport entity. For bidirectional switching, automatic protection switching (APS) protocol is required to coordinate the two endpoints so that both have the same bridge and selector settings, even for a unidirectional failure. This type is applicable for 1+1 and 1:1 protection.

2.3. Protection operation type

The linear protection operation types can be a non-revertive operation type or a revertive operation type.

- o Non-revertive operation: The normal traffic will not be switched back to the working transport entity even after a protection switching cause has cleared. This is generally accomplished by replacing the previous switch request with a "Do not Revert (DNR)" request, which has a low priority.
- o Revertive operation: The normal traffic is restored to the working transport entity after the condition(s) causing the protection switching has cleared. In the case of clearing a command (e.g., Forced Switch), this happens immediately. In the case of clearing of a defect, this generally happens after the expiry of a "Wait-to-Restore (WTR)" timer, which is used to avoid chattering of selectors in the case of intermittent defects.

3. Protection switching trigger conditions

3.1. Fault conditions

Fault conditions mean the requests generated by the local OAM function.

- o Signal Failure (SF): If an endpoint detects a failure by OAM function or other mechanism, it will submit a local signal failure (local SF) to APS module to request a protection switching. The local SF could be on working transport entity or protection transport entity.
- o Signal Degrade (SD): If an endpoint detects signal degrade by OAM function or other mechanism, it will submit a local signal failure (local SD) to APS module to request a protection switching. The local SD could be on working transport entity or protection transport entity.

3.2. External commands

The external command issues an appropriate external request on to the protection process.

3.2.1. End-to-end commands

These commands are applied to both local and remote nodes. When the APS protocol is present, these commands are signaled to the far end of the connection. In bidirectional switching, these commands affect the bridge and selector at both ends.

- o Lockout of Protection (LO): This command is used to provide operator a tool for temporarily disabling access to the protection transport entity.
- o Manual switch (MS): This command is used to provide operator a tool for temporarily switching normal traffic to working transport entity (MS-W) or protection transport entity (MS-P), unless a higher priority switch request (i.e., LP, FS, or SF) is in effect.
- o Forced switch (FS): This command is used to provide operator a tool for temporarily switching normal traffic from working transport entity to protection transport entity, unless a higher priority switch request (i.e., LP) is in effect.
- o Exercise (EXER): Exercise is a command to test if the APS communication is operating correctly. The EXER command will not affect the state of the protection selector and bridge.

- o Clear: This command between management and local protection process is not a request sent by APS to other endpoints. It is used to clear the active near end external command or WTR state.

3.2.2. Local commands

These commands apply only to the near end (local node) of the protection group. Even when an APS protocol is supported, they are not signalled to the far end.

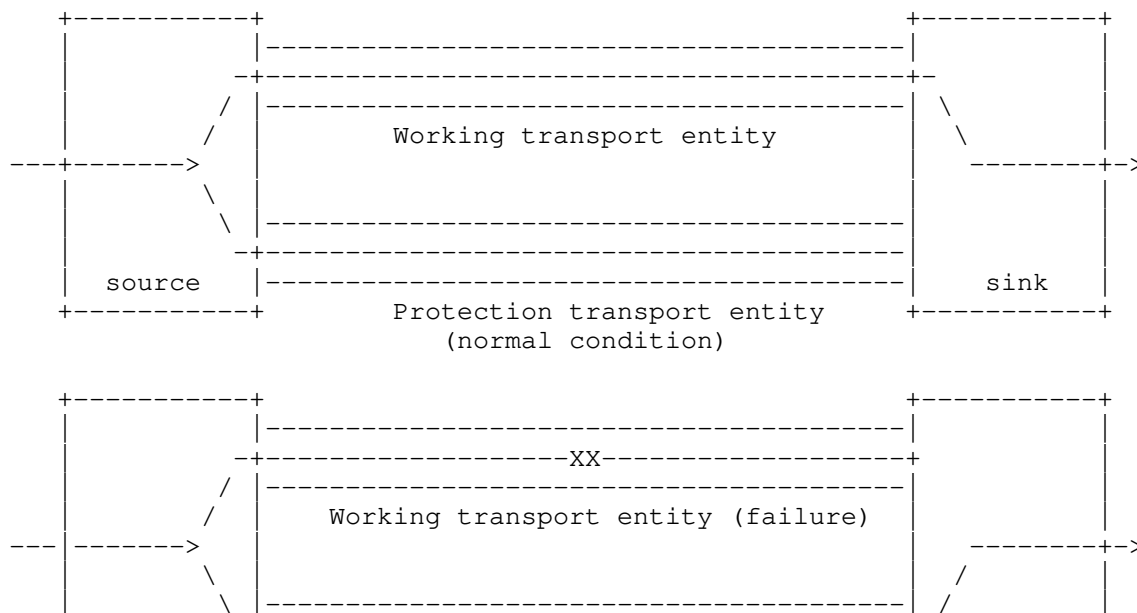
- o Freeze: This command freezes the state of the protection group. Until the freeze is cleared, additional near end commands are rejected and condition changes and received APS information are ignored. When the Freeze command is cleared, the state of the protection group is recomputed based on the condition and received APS information.

Because the freeze is local, if the freeze is issued at one end only, a failure of protocol can occur as the other end is open to accept any operator command or a fault condition.

- o Clear Freeze: This command clears the local freeze.

4. Protection switching schemes

4.1. 1+1 unidirectional protection switching



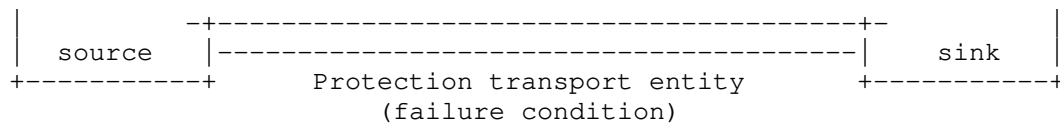
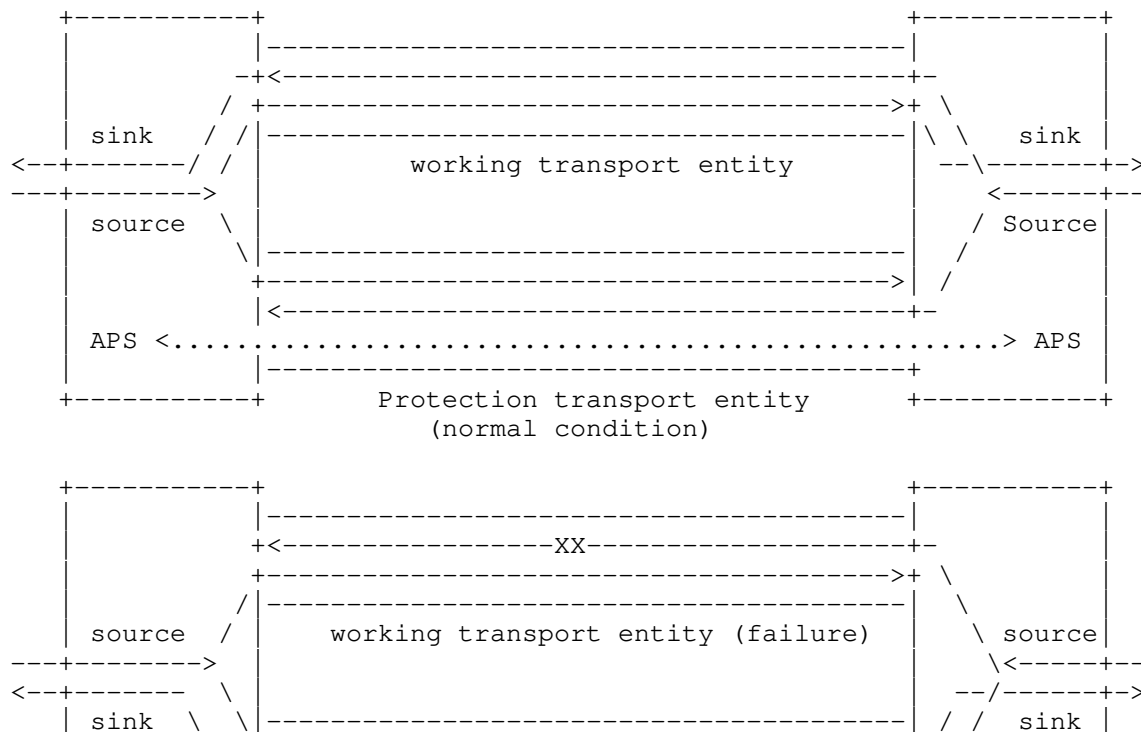


Figure 1: 1+1 unidirectional linear protection switching

1+1 unidirectional protection switching is the simplest protection switching mechanism. The normal traffic is permanently bridged on both the working and protection transport entities at the source endpoint of the protection domain. In normal condition, the sink endpoint receives traffic from working transport entity. If the sink endpoint detects a failure on working transport entity, it will switch to receive traffic from protection transport entity. 1+1 unidirectional protection switching is recommended to be used for unidirectional transport entity.

Note that 1+1 unidirectional protection switching does not need APS coordination protocol since it only perform protection switching based on the local request.

4.2. 1+1 bidirectional protection switching



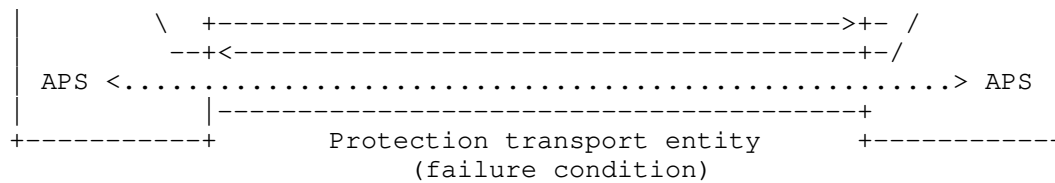


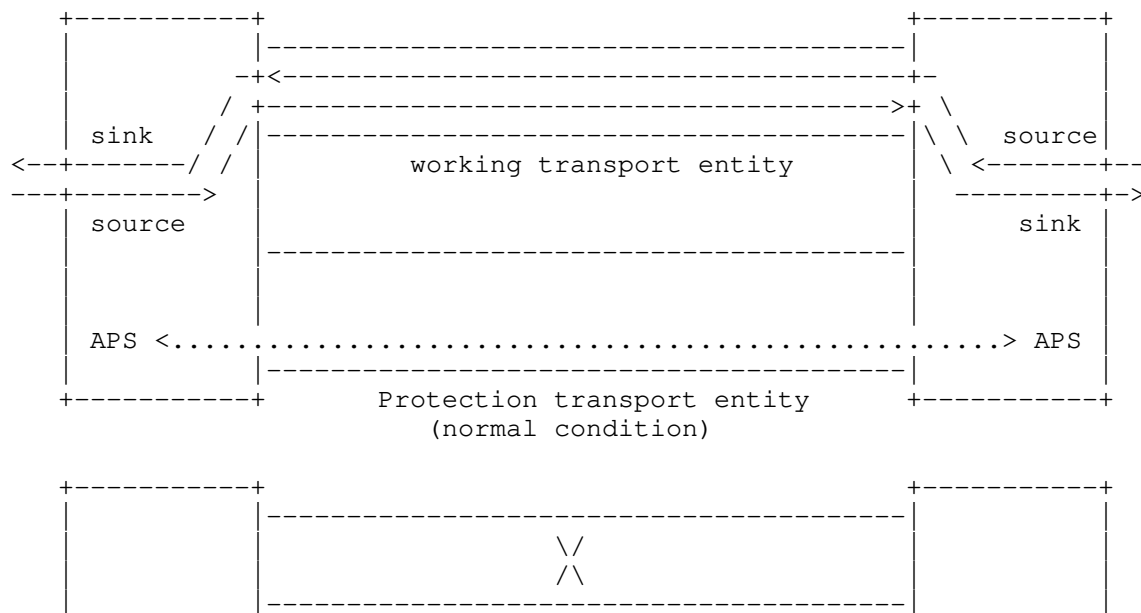
Figure 2: 1+1 bidirectional linear protection switching

In 1+1 bidirectional protection switching, for each direction, the normal traffic is permanently bridged on both the working and protection transport entities at the source endpoint of the protection domain. In normal condition, for each direction, the sink endpoint receives traffic from working transport entity.

If the sink endpoint detects a failure on the working transport entity, it will switch to receive traffic from protection transport entity. It will also send an APS message to inform the sink endpoint on another direction to switch to receive traffic from protection transport entity.

APS mechanism is necessary to coordinate the two endpoints of transport entity and implement 1+1 bidirectional protection switching even for a unidirectional failure.

4.3. 1:1 bidirectional protection switching



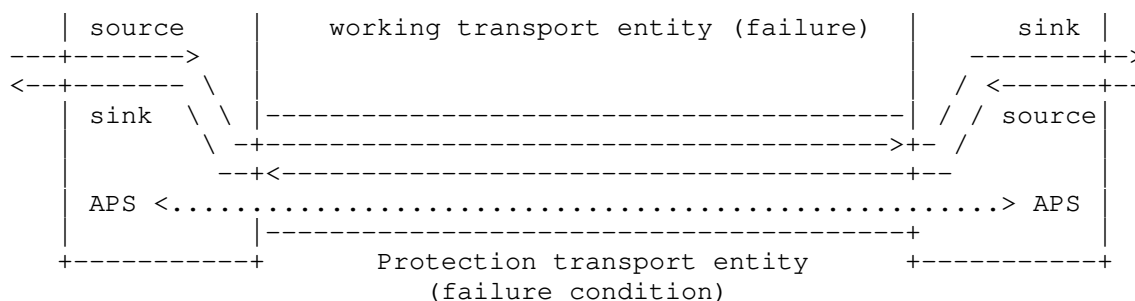


Figure 3: 1:1 bidirectional linear protection switching

In 1:1 bidirectional protection switching, for each direction, the source endpoint sends traffic on either working transport entity or protection transport entity. The sink endpoint receives the traffic from the transport entity where the source endpoint sends on.

In normal condition, for each direction, the source endpoint and sink endpoint send and receive traffic from working transport entity.

If the sink endpoint detects a failure on the working transport entity, it will switch to send and receive traffic from protection transport entity. It will also send an APS message to inform the sink endpoint on another direction to switch to send and receive traffic from protection transport entity.

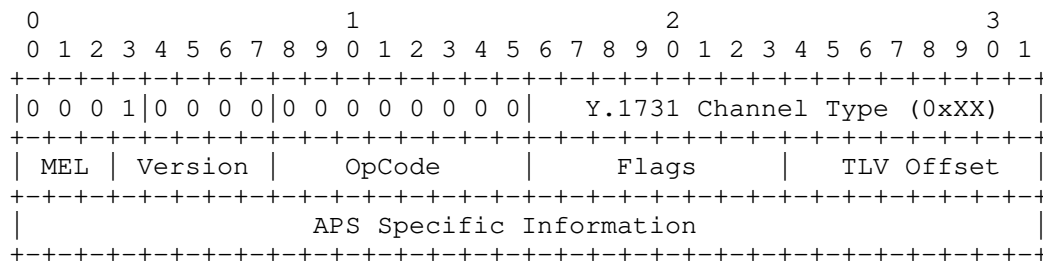
APS mechanism is necessary to coordinate the two endpoints of transport entity and implement 1:1 bidirectional protection switching even for a unidirectional failure.

5. APS protocol

5.1. APS PDU format

APS packets MUST be sent over a G-ACh as defined in [RFC5586].

The format of APS PDU is specified in Figure 4 below.



```

|      End TLV      |
+---+---+---+---+

```

Figure 4: APS PDU format

The following values shall be used for APS PDU:

- o The Y.1731 Channel Type is set as defined in [BHH_MPLS-TP_OAM]
- o MEL: set as defined in [BHH_MPLS-TP_OAM];
- o Version: 0x00
- o OpCode: 0d39 (=0x27)
- o Flags: 0x00
- o TLV Offset: 4
- o End TLV: 0x00

The format of the APS-specific information is defined in Figure 5

| 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 | | | | | | | | |
| Request | | | | | | | | Pr.Type | | | | | | | | Requested | | | | | | | | Bridged | | | | | | | | Reserved(0) | | | | | | | |
| / | | | | | | | | -+--+ | | | | | | | | | | | | | | | | | | | | | | | | T | | | | | | | |
| State | | | | | | | | A B D R | | | | | | | | Signal | | | | | | | | Signal | | | | | | | | | | | | | | | |

Figure 5: APS specific information format

All bits defined as "Reserved" shall be transmitted as 0 and ignored on reception.

- o Request/State:

The 4 bits indicate the protection switching request type. See Figure 6 for the code of each request/state type.

In case that there are multiple protection switching requests, only the protection switching request with the highest priority will be processed.

| | |
|---------------|---------------|
| Request/State | code/priority |
|---------------|---------------|

| | |
|-----------------------------------|----------------|
| Lockout of Protection (LO) | 1111 (highest) |
| Signal Fail for Protection (SF-P) | 1110 |
| Forced Switch (FS) | 1101 |
| Signal Fail for Working (SF-W) | 1011 |
| Signal Degrade | 1001 |
| Manual Switch | 0111 |
| Wait to Restore (WTR) | 0101 |
| Exercise (EXER) | 0100 |
| Reverse Request (RR) | 0010 |
| Do Not Revert (DNR) | 0001 |
| No Request (NR) | 0000 (lowest) |

Figure 6: Protection switching request code/priority

- o Protection type (Pr.Type):

The 4 bits are used to specify the protection type.

A: reserved (set by default to 1)
 B: 0 - 1+1 (permanent bridge)
 1 - 1:1 (no permanent bridge)
 D: 0 - Unidirectional switching
 1 - Bidirectional switching
 R: 0 - Non-revertive operation
 1 - Revertive operation

- o Requested signal:

This byte is used to indicate the traffic that the near end requests to be carried over the protection entity.

value = 0 Null traffic
 value = 1 Normal traffic 1
 value = 2~255 Reserved

- o Bridged signal:

This byte is used to indicate the traffic that is bridged onto the protection entity.

value = 0 Null traffic
value = 1 Normal traffic 1
value = 2~255 Reserved

o Bridge Type (T):

This bit is used to further specify the type of non-permanent bridge for 1:1 protection switching.

value = 0 Selector bridge
value = 1 Broadcast bridge

o Reserved:

This field should be set to zero.

5.2. APS transmission

The APS message should be transported on protection transport entity by encapsulated with the protection transport entity label. If an endpoint receives APS-specific information from the working entity, it should ignore this information, and should detect the Failure of Protocol defect (see Section 6).

A new APS packet must be transmitted immediately when a change in the transmitted status occurs. The first three APS packets should be transmitted as fast as possible only if the APS information to be transmitted has been changed so that fast protection switching is possible even if one or two APS packets are lost or corrupted. The interval of the first three APS packets should be 3.3ms. APS packets after the first three should be transmitted with the interval of 5 seconds.

If no valid APS-specific information is received, the last valid received information remains applicable.

5.3. Hold-off timer

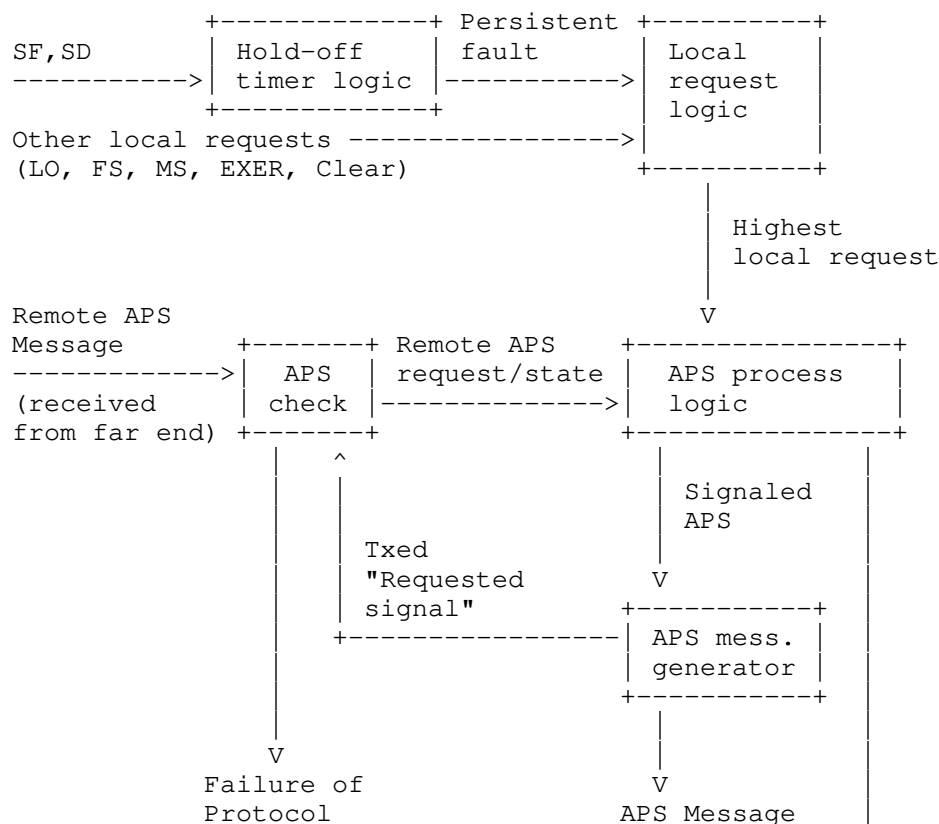
In order to coordinate timing of protection switches at multiple layers, a hold-off timer may be required. The purpose is to allow a server layer protection switch to have a chance to fix the problem before switching at a client layer.

Each protection group should have a provisioned hold-off timer. The suggested range of the hold-off timer is 0 to 10 seconds in steps of 100 ms (accuracy of +/-5 ms).

When a new defect or more severe defect occurs (new SF/SD) on the transport entity that currently carries traffic, this event will not be reported immediately to protection switching if the provisioned hold-off timer value is non-zero. Instead, the hold-off timer will be started. When the hold-off timer expires, it will be checked whether a defect still exists on the transport entity that started the timer. If it does, that defect will be reported to protection switching. The defect need not be the same one that started the timer.

This hold-off timer mechanism shall be applied for both working and protection transport entities.

6. Protection switching logic



Detection

V
Set local
bridge/selector

Figure 7: Protection Switching Logic

Figure 7 describes the protection switching logic.

One or more local protection switching requests may be active. The "local request logic" determines which of these requests is highest using the order of priority given in Figure 6. This highest local request information is passed on to the "APS process logic". Note that an accepted Clear command, clearance of SF(-P) or expiration of WTR timer shall not be processed by the local request logic, but shall be considered as the highest local request and submitted to the APS process logic for processing.

The remote APS message is received from the far end and is subjected to the validity check and mismatch detection in "APS check". Failure of Protocol situations are as follows:

- o The "B" field mismatch due to incompatible provisioning;
- o The reception of APS message from the working entity due to working/protection configuration mismatch;
- o No match in sent "Requested traffic" and received "requested signal" for more than 50 ms;
- o No APS message is received on the protection transport entity during at least 3.5 times the long APS interval (e.g. at least 17.5 seconds) and there is no defect on the protection transport entity.

Provided the "B" field matches:

- o If "D" bit mismatches, the bidirectional side will fall back to unidirectional switching.
- o If the "R" bit mismatches, one side will clear switches to "WTR" and the other will clear to "DNR". The two sides will interwork and the traffic is protected.
- o If the "T" bit mismatches, the side using a broadcast bridge will fall back to using a selector bridge.

The APS message with invalid information should be ignored, and the last valid received information remains applicable.

The linear protection switching algorithm commences immediately every time one of the input signals changes, i.e., when the status of any local request changes, or when a different APS specific information is received from the far end. The consequent actions of the algorithm are also initiated immediately, i.e., change the local bridge/selector position (if necessary), transmit a new APS specific information (if necessary), or detect the failure of protocol defect if the protection switching is not completed within 50 ms.

The state transition is calculated in the "APS process logic" based on the highest local request, the request of the last received "Request/State" information, and state transition tables defined in Section 7, as follows:

- o If the highest local request is Clear, clearance of SF(-P) or of SD, or expiration of WTR, a state transition is calculated first based on the highest local request and state machine table for local requests to obtain an intermediate state. This intermediate state is the final state in case of clearance of SF-P otherwise, starting at this intermediate state, the last received far end request and the state machine table for far end requests are used to calculate the final state.
- o If the highest local request is neither Clear, nor clearance of SF(-P) or of SD, nor expiration of WTR, the APS process logic compares the highest local request with the request of the last received "Request/State" information based on Figure 6.
 - 1. If the highest local request has higher or equal priority, it is used with the state transition table for local requests defined in Section 7 to determine the final state; otherwise
 - 2. The request of the last received "Request/State" information is used with the state transition table for far end requests defined in Annex A to determine the final state.

The "APS message generator" generates APS specific information with the signaled APS information for the final state from the state transition calculation (with coding as described in Figure 5).

7. Protection switching state transition table

In this section, state transition tables for the following protection switching configurations are described.

- o 1:1 bidirectional (revertive mode, non-revertive mode);
- o 1+1 bidirectional (revertive mode, non-revertive mode);

- o 1+1 unidirectional (revertive mode, non-revertive mode).

Note that any other global or local request which is not described in state transition tables does not trigger any state transition.

The states specified in the state transition tables can be described as follows:

- o No request: No Request is the state entered by the local priority under all conditions where no local protection switching requests (including wait-to-restore and do-not-revert) are active. NR can also indicate that the highest local request is overridden by the far end request, whose priority is higher than the highest local request. Normal traffic signal is selected from the corresponding transport entity.
- o Lockout, Signal Fail(P): The access by the normal traffic to the protection transport entity is NOT allowed, due to the SF detected on the protection entity or due to the lockout of protection command applied. The normal traffic is carried by the working transport entity, regardless of the fault/degrade condition possibly present (due to the highest priority of the switching triggers leading to this state).
- o Forced Switch, Signal Fail(W), Signal Degrade(W), Signal Degrade(P), Manual Switch: A switching trigger, NOT resulting in the protection transport entity unavailability is present. The normal traffic is selected either from the corresponding working transport entity or from the protection transport entity, according to the behaviour of the specific switching trigger.
- o Wait to Restore: In revertive operation, after the clearing of an SF or SD on working transport entity, maintains normal traffic as selected from the protection transport entity until a wait-to-restore timer expires or another request with higher priority, including a clear command, is received. This is used to prevent frequent operation of the selector in the case of intermittent failures.
- o Do not revert: In non-revertive operation, this is used to maintain a normal traffic to be selected from the protection transport entity.
- o Exercise: Exercise of the APS protocol.
- o Reverse Request: The near end will enter and signal Reverse Request only in response to an EXER from the far end.

[State transition tables are shown at the end of the PDF form of this document.]

8. Security considerations

To be added in a future version of the document.

9. IANA considerations

To be added in a future version of the document.

10. Acknowledgements

The authors would like to thank Hao Long, Vincenzo Sestito, Italo Busi, Igor Umansky for their input to and review of the current document.

11. References

11.1. Normative References

- [RFC5317] Bryant, S. and L. Andersson, "Joint Working Team (JWT) Report on MPLS Architectural Considerations for a Transport Profile", RFC 5317, February 2009.
- [RFC5586] Bocci, M., Vigoureux, M., and S. Bryant, "MPLS Generic Associated Channel", RFC 5586, June 2009.
- [RFC5654] Niven-Jenkins, B., Brungard, D., Betts, M., Sprecher, N., and S. Ueno, "Requirements of an MPLS Transport Profile", RFC 5654, September 2009.
- [BHH_MPLS-TP_OAM] Busi, I., van Helvoort, H., and J. He, "MPLS-TP OAM based on Y.1731", draft-bhh-mpls-tp-oam-y1731-07, July 2011.

11.2. Informative References

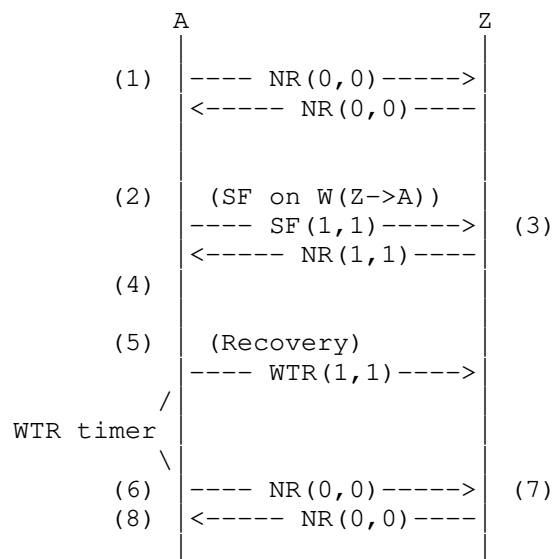
- [RFC6372] Sprecher, N. and A. Farrel, "MPLS-TP Survivability Framework", RFC 6372, Sept 2011.

Appendix A. Operation examples of APS protocol

The sequence diagrams shown in this section are only a few examples of the APS operations. The first APS message which differs from the previous APS message is shown. The operation of hold-off timer is omitted. The fields whose values are changed during APS packet exchange are shown in the APS packet exchange. They are Request/

State, requested traffic, and bridged traffic. For an example, SF(0,1) represents an APS packet with the following field values: Request/State = SF, requested signal = 0, and bridged signal = 1. The values of the other fields remain unchanged from the initial configuration. The signal numbers 0 and 1 refer to null signal and normal traffic signal, respectively. W(A->Z) and P(A->Z) indicate the working and protection paths in the direction of A to Z, respectively.

Example 1. 1:1 bidirectional protection switching (revertive mode) - Unidirectional SF case



(1) The protection domain is operating without any defect, and the working entity is used for delivering the normal traffic.

(2) Signal Fail occurs on the working entity in the Z to A direction. Selector and bridge of node A select protection entity. Node A generates SF(r=1, b=1) message.

(3) Upon receiving SF(r=1, b=1), node Z sets selector and bridge to protection entity. As there is no local request in node Z, node Z generates NR(r=1, b=1) message.

(4) Node A confirms that the far end is also selecting protection entity.

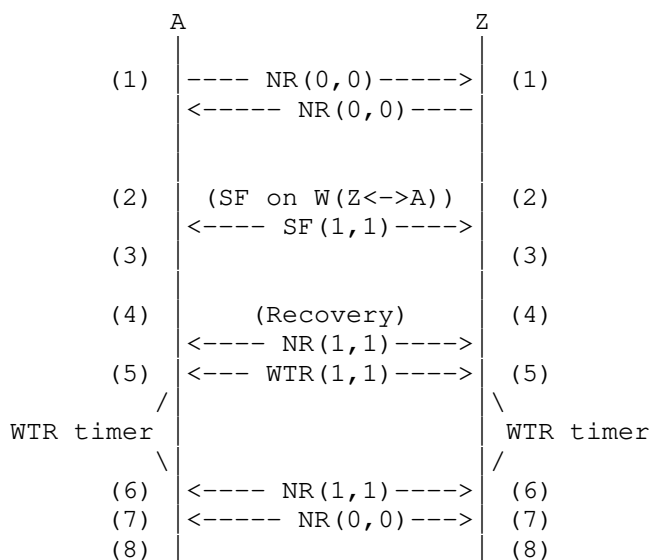
(5) Node A detects clearing of SF condition, starts the WTR timer, and sends WTR(r=1, b=1) message.

(6) At expiration of the WTR timer, node A sets selector and bridge to working entity and sends NR(r=0, b=0) message.

(7) Node Z is notified that the far end request has been cleared, and sets selector and bridge to working entity.

(8) It is confirmed that the far end is also selecting working entity.

Example 2. 1:1 bidirectional protection switching (revertive mode) - Bidirectional SF case



(1) The protection domain is operating without any defect, and the working entity is used for delivering the normal traffic.

(2) Nodes A and Z detect local Signal Fail conditions on the working entity, set selector and bridge to protection entity, and generate SF(r=1, b=1) messages.

(3) Upon receiving SF(r=1, b=1), each node confirms that the far end is also selecting protection entity.

(4) Each node detects clearing of SF condition, and sends NR(r=1, b=1) message as the last received APS message was SF.

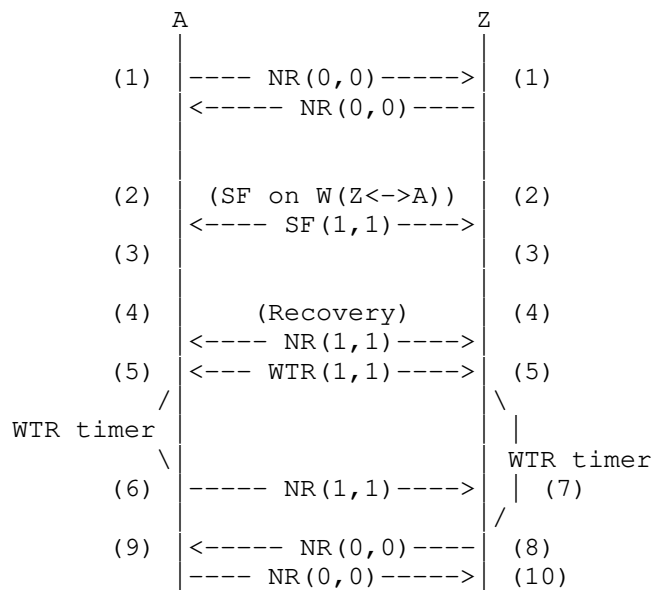
(5) Upon receiving NR(r=1, b=1), each node starts the WTR timer and sends WTR(r=1, b=1).

(6) At expiration of the WTR timer, each node sends NR(r=1, b=1) as the last received APS message was WTR.

(7) Upon receiving NR(r=1, b=1), each node sets selector and bridge to working entity and sends NR(r=0, b=0) message.

(8) It is confirmed that the far end is also selecting working entity.

Example 3. 1:1 bidirectional protection switching (revertive mode) - Bidirectional SF case - Inconsistent WTR timers



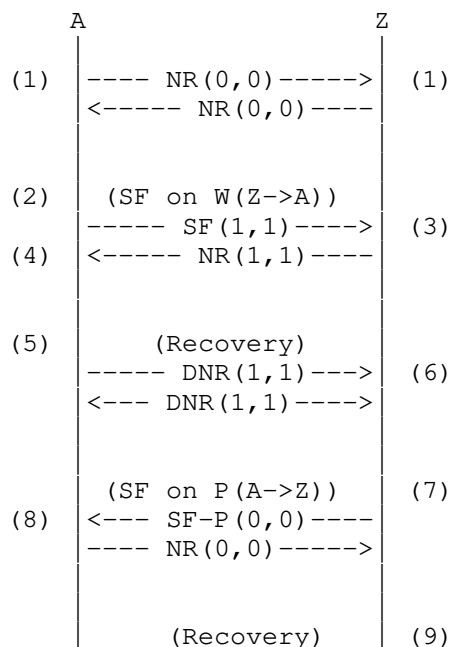
(1) The protection domain is operating without any defect, and the working entity is used for delivering the normal traffic.

(2) Nodes A and Z detect local Signal Fail conditions on the working entity, set selector and bridge to protection entity, and generate SF(r=1, b=1) messages.

(3) Upon receiving SF(r=1, b=1), each node confirms that the far end is also selecting protection entity.

- (4) Each node detects clearing of SF condition, and sends NR(r=1, b=1) message as the last received APS message was SF.
- (5) Upon receiving NR(r=1, b=1), each node starts the WTR timer and sends WTR(r=1, b=1).
- (6) At expiration of the WTR timer in node A, node A sends NR(r=1, b=1) as the last received APS message was WTR.
- (7) At node Z, the received NR(r=1, b=1) is ignored as the local WTR has a higher priority.
- (8) At expiration of the WTR timer in node Z, node Z node sets selector and bridge to working entity, and sends NR(r=0, b=0) message.
- (9) Upon receiving NR(r=0, b=0), node A sets selector and bridge to working entity and sends NR(r=0, b=0) message.
- (10) It is confirmed that the far end is also selecting working entity.

Example 4. 1:1 bidirectional protection switching (non-revertive mode) - Unidirectional SF on working followed by unidirectional SF on protection



```

      |<----- NR(0,0)-----|

```

(1) The protection domain is operating without any defect, and the working entity is used for delivering the normal traffic.

(2) Signal Fail occurs on the working entity in the Z to A direction. Selector and bridge of node A select the protection entity. Node A generates SF(r=1, b=1) message.

(3) Upon receiving SF(r=1, b=1), node Z sets selector and bridge to protection entity. As there is no local request in node Z, node Z generates NR(r=1, b=1) message.

(4) Node A confirms that the far end is also selecting protection entity.

(5) Node A detects clearing of SF condition, and sends DNR(r=1, b=1) message.

(6) Upon receiving DNR(r=1, b=1), node Z also generates DNR(r=1, b=1) message.

(7) Signal Fail occurs on the protection entity in the A to Z direction. Selector and bridge of node Z select the working entity. Node Z generates SF-P(r=0, b=0) message.

(8) Upon receiving SF-P(r=0, b=0), node A sets selector and bridge to working entity, and generates NR(r=0, b=0) message.

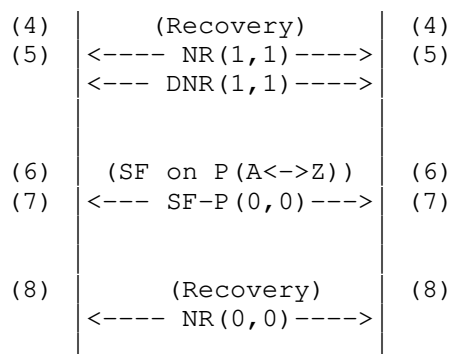
(9) Node Z detects clearing of SF condition, and sends NR(r=0, b=0) message.

Exmaple 5. 1:1 bidirectional protection switching (non-revertive mode) - Bidirectional SF on working followed by bidirectional SF on protection

```

      A                                     Z
      |                                     |
(1)  |----- NR(0,0)----->          | (1)
      |<----- NR(0,0)-----|
      |
(2)  | (SF on W(A<->Z)) |          | (2)
(3)  |<----- SF(1,1)----->          | (3)
      |                                     |

```



(1) The protection domain is operating without any defect, and the working entity is used for delivering the normal traffic.

(2) Nodes A and Z detect local Signal Fail conditions on the working entity, set selector and bridge to protection entity, and generate SF(r=1, b=1) messages.

(3) Upon receiving SF(r=1, b=1), each node confirms that the far end is also selecting protection entity.

(4) Each node detects clearing of SF condition, and sends NR(r=1, b=1) message as the last received APS message was SF.

(5) Upon receiving NR(r=1, b=1), each node sends DNR(r=1, b=1).

(6) Signal Fail occurs on the protection entity in both directions. Selector and bridge of each node selects the working entity. Each node generates SF-P(r=0, b=0) message.

(7) Upon receiving SF-P(r=0, b=0), each node confirms that the far end is also selecting working entity

(8) Each node detects clearing of SF condition, and sends NR(r=0, b=0) message.

Authors' Addresses

Huub van Helvoort (editor)
Huawei Technologies

Email: huub.van.helvoort@huawei.com

Jeong-dong Ryoo (editor)
ETRI

Email: ryoo@etri.re.kr

Haiyan Zhang
Huawei Technologies

Email: zhanghaiyan@huawei.com

Feng Huang
Alcatel-Lucent Shanghai Bell

Email: feng.f.huang@alcatel-sbell.com.cn

Han Li
China Mobile

Email: lihan@chinamobile.com

Alessandro D'Alessandro
Telecom Italia

Email: alessandro.dalessandro@telecomitalia.it