

Internet Engineering Task Force
Internet-Draft
Updates: 4379,6424 (if approved)
Intended status: Standards Track
Expires: December 15, 2014

N. Akiya
G. Swallow
Cisco Systems
S. Litkowski
B. Decraene
Orange
J. Drake
Juniper Networks
June 13, 2014

Label Switched Path (LSP) Ping/Trace Multipath Support for
Link Aggregation Group (LAG) Interfaces
draft-akiya-mpls-lsp-ping-lag-multipath-00

Abstract

This document defines an extension to the Multiprotocol Label Switching (MPLS) Label Switched Path (LSP) Ping and Traceroute to describe Multipath Information for Link Aggregation (LAG) member links separately, thus allowing MPLS LSP Ping and Traceroute to discover and exercise specific paths of layer 2 Equal-Cost Multipath (ECMP) over LAG interfaces.

This document updates RFC4379 and RFC6424.

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 December 15, 2014.

Copyright Notice

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

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

Table of Contents

1. Introduction	3
1.1. Terminology	3
1.2. Background	3
2. Overview	4
3. Mechanism to Discover L2 ECMP Multipath	5
4. Mechanism to Validate L2 ECMP Traversal	6
5. LAG Interface Info TLV	7
6. DDMAP TLV DS Flags: G	8
7. Interface Index Sub-TLV	9
8. Detailed Interface and Label Stack TLV	10
8.1. Sub-TLVs	11
8.1.1. Incoming Label Stack Sub-TLV	12
8.1.2. Incoming Interface Index Sub-TLV	12
9. Security Considerations	13
10. IANA Considerations	13
10.1. LAG Interface Info TLV	13
10.2. Interface Index Sub-TLV	13
10.3. Detailed Interface and Label Stack TLV	14
10.4. New Sub-Registry	14
10.4.1. DS Flags	14
10.4.2. Sub-TLVs for TLV Type TBD3	15
11. Acknowledgements	15
12. References	15
12.1. Normative References	15
12.2. Informative References	15
Appendix A. LAG with L2 Switch Issues	16
A.1. Equal Numbers of LAG Members	16
A.2. Deviating Numbers of LAG Members	17
A.3. LAG Only on Right	17
A.4. LAG Only on Left	17
Authors' Addresses	17

1. Introduction

1.1. Terminology

The following acronyms/terminologies are used in this document:

- o MPLS - Multiprotocol Label Switching.
- o LSP - Label Switched Path.
- o LSR - Label Switching Router.
- o ECMP - Equal-Cost Multipath.
- o LAG - Link Aggregation.
- o Initiating LSR - LSR which sends MPLS echo request.
- o Responder LSR - LSR which receives MPLS echo request and sends MPLS echo reply.

1.2. Background

The Multiprotocol Label Switching (MPLS) Label Switched Path (LSP) Ping and Traceroute [RFC4379] are powerful tools designed to diagnose all available layer 3 paths of LSPs, i.e. provides diagnostic coverage of layer 3 Equal-Cost Multipath (ECMP). In many MPLS networks, Link Aggregation (LAG) as defined in [IEEE802.1AX], which provide layer 2 ECMP, are often used for various reasons. MPLS LSP Ping and Traceroute tools were not designed to discover and exercise specific paths of layer 2 ECMP. Result raises a limitation for following scenario when LSP X traverses over LAG Y:

- o MPLS switching of LSP X over one or more member links of LAG Y is succeeding.
- o MPLS switching of LSP X over one or more member links of LAG Y is failing.
- o MPLS echo request for LSP X over LAG Y is load balanced over a member link which is MPLS switching successfully.

With above scenario, MPLS LSP Ping and Traceroute will not be able to detect the MPLS switching failure of problematic member link(s) of the LAG. In other words, lack of layer 2 ECMP discovery and exercise capability can produce an outcome where MPLS LSP Ping and Traceroute can be blind to MPLS switching failures over LAG interface that are impacting MPLS traffic. It is, thus, desirable to extend the MPLS

LSP Ping and Traceroute to have deterministic diagnostic coverage of LAG interfaces.

2. Overview

This document defines an extension to the MPLS LSP Ping and Traceroute to describe Multipath Information for LAG member links separately, thus allowing MPLS LSP Ping and Traceroute to discover and exercise specific paths of layer 2 ECMP over LAG interfaces. Reader is expected to be familiar with mechanics of the MPLS LSP Ping and Traceroute described in Section 3.3 of [RFC4379] and Downstream Detailed Mapping TLV (DDMAP) described in Section 3.3 of [RFC6424].

MPLS echo request carries a DDMAP and an optional TLV to indicate that separate load balancing information for each layer 2 nexthop over LAG is desired in MPLS echo reply. Responder LSR places the same optional TLV in the MPLS echo reply to provide acknowledgement back to the initiator. It also adds, for each downstream LAG member, a load balance information (i.e. multipath information and interface index). For example:

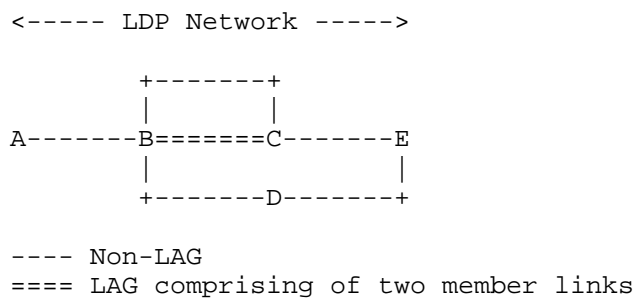


Figure 1: Example LDP Network

When node A is initiating LSP Traceroute to node E, node B will return to node A load balance information for following entries.

1. Downstream C over Non-LAG (upper path).
2. First Downstream C over LAG (middle path).
3. Second Downstream C over LAG (middle path).
4. Downstream D over Non-LAG (lower path).

This document defines:

- o In Section 3, a mechanism to discover L2 ECMP multipath information;
- o In Section 4, a mechanism to validate L2 ECMP traversal in some LAG provisioning models;
- o In Section 5, the LAG Interface Info TLV;
- o In Section 6, the LAG Description Indicator flag;
- o In Section 7, the Interface Index Sub-TLV;
- o In Section 8, the Detailed Interface and Label Stack TLV.

3. Mechanism to Discover L2 ECMP Multipath

The MPLS echo request carries a DDMAP and the LAG Interface Info TLV (described in Section 5) to indicate that separate load balancing information for each layer 2 nexthop over LAG is desired in MPLS echo reply. Responder LSR:

- o MUST add the LAG Interface Info TLV in the MPLS echo reply to provide acknowledgement back to the initiator. Downstream LAG Info Accommodation flag MUST be set in LAG Interface Info Flags.
- o For each downstream that is a LAG interface:
 - * MUST add DDMAP in the MPLS echo reply.
 - * MUST set LAG Description Indicator flag in DS Flags (described in Section 6) of DDMAP.
 - * All fields and Sub-TLVs, except for Multipath Data Sub-TLV and Interface Index Sub-TLV, are set/added to DDMAP to describe this LAG interface, as per [RFC6424].
 - * For each LAG member link of this LAG interface:
 - + MUST add Interface Index Sub-TLV (described in Section 7) with LAG Member Link Indicator flag set in Interface Index Flags, describing this LAG member link.
 - + MUST add Multipath Data Sub-TLV for this LAG member link, if received DDMAP requested multipath information.

Each LAG member link is described with Interface Index Sub-TLV and conditionally with Multipath Data Sub-TLV (if multipath information is requested). If both Sub-TLVs are placed in the DDMAP to describe

a LAG member link, Interface Index Sub-TLV MUST be added first with Multipath Data Sub-TLV immediately following.

For example, a responder LSR possessing a LAG interface with two member links would send the following DDMAP for this LAG interface:

```

      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
+-----+-----+-----+-----+-----+-----+-----+-----+
| DDMAP fields describing LAG interface with DS Flags G set |
+-----+-----+-----+-----+-----+-----+-----+-----+
|      Interface Index Sub-TLV of LAG member link #1      |
+-----+-----+-----+-----+-----+-----+-----+-----+
|      Multipath Data Sub-TLV LAG member link #1          |
+-----+-----+-----+-----+-----+-----+-----+-----+
|      Interface Index Sub-TLV of LAG member link #2      |
+-----+-----+-----+-----+-----+-----+-----+-----+
|      Multipath Data Sub-TLV LAG member link #2          |
+-----+-----+-----+-----+-----+-----+-----+-----+
|      Label Stack Sub-TLV                                |
+-----+-----+-----+-----+-----+-----+-----+-----+

```

Figure 2: LAG Interface DDMAP Example

These procedures allow initiating LSR to:

- o Identify whether responder LSR understands this mechanism.
- o Identify whether each DDMAP describes a LAG interface or a non-LAG interface.
- o Obtain multipath information which is expected to traverse the specific LAG member link described by interface index.

4. Mechanism to Validate L2 ECMP Traversal

The MPLS echo request is sent with a DDMAP with DS Flags I set and the optional LAG Interface Info TLV to indicate the request for Detailed Interface and Label Stack TLV with additional LAG member link information (i.e. interface index) in the MPLS echo reply. Responder LSR MUST:

- o Add LAG Interface Info TLV in the MPLS echo reply to provide acknowledgement back to the initiator. Upstream LAG Info Accommodation flag MUST be set in LAG Interface Info Flags.
- o Add the Detailed Interface and Label Stack TLV (described in Section 8) in the MPLS echo reply.

- o Add the Incoming Interface Index Sub-TLV (described in Section 8.1.2) for LAG interfaces. The LAG Member Link Indicator flag MUST be set in Interface Index Flags, and the incoming Interface Index set to LAG member link which received the MPLS echo request.

Described procedures allow initiating LSR to know:

- o The expected load balance information of every LAG member link, at LSR with TTL=n.
- o The actual incoming interface at LSR with TTL=n+1, including the interface index of LAG member link if incoming interface is a LAG interface.

Note that defined procedures will provide a deterministic result for LAG interfaces that are back-to-back connected between routers (i.e. no L2 switch in between). If there is a L2 switch between LSR at TTL=n and LSR at TTL=n+1, there is no guarantee that traversal of every LAG member link at TTL=n will result in reaching different interface index at TTL=n+1. Issues resulting from LAG with L2 switch in between are further described in Appendix A. LAG provisioning models in operated network should be considered when analyzing the output of LSP Traceroute exercising L2 ECMPs.

5. LAG Interface Info TLV

The LAG Interface Info object is a new TLV that MAY be included in the MPLS echo request message. An MPLS echo request MUST NOT include more than one LAG Interface Info object. Presence of LAG Interface Info object is a request that responder LSR describes upstream and downstream LAG interfaces according to procedures defined in this document. If the responder LSR is able to accommodate this request, then the LAG Interface Info object MUST be included in the MPLS echo reply message.

LAG Interface Info TLV Type is TBD1. Length is 4. The Value field of LAG Interface TLV has following format:

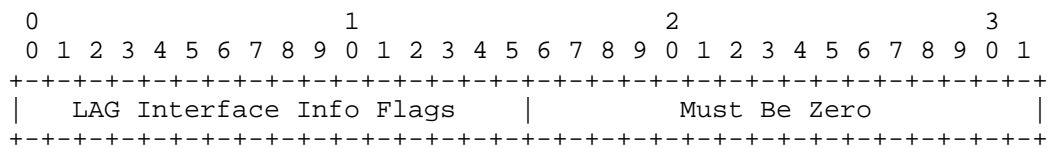


Figure 3: LAG Interface Info TLV

LAG Interface Info Flags

LAG Interface Info Flags field is a bit vector with following format.

```

      0                                     1
    0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5
+-----+-----+-----+-----+
| Must Be Zero (Reserved) |U|D|
+-----+-----+-----+-----+

```

Two flags are defined: U and D. The remaining flags MUST be set to zero when sending and ignored on receipt. Both U and D flags MUST be cleared in MPLS echo request message when sending, and ignored on receipt. Either or both U and D flags MAY be set in MPLS echo reply message.

Flag	Name and Meaning
U	Upstream LAG Info Accommodation
D	Downstream LAG Info Accommodation

When this flag is set, LSR is capable of placing Incoming Interface Index Sub-TLV, describing LAG member link, in the Detailed Interface and Label Stack TLV.

When this flag is set, LSR is capable of placing Interface Index Sub-TLV and Multipath Data Sub-TLV, describing LAG member link, in the Downstream Detailed Mapping TLV.

6. DDMAP TLV DS Flags: G

One flag, G, is added in DS Flags field of the DDMAP TLV. In the MPLS echo request message, G flag MUST be cleared when sending, and ignored on receipt. In the MPLS echo reply message, G flag MUST be set if the DDMAP TLV describes a LAG interface. It MUST be cleared otherwise.

DS Flags

DS Flags G is added, in Bit Number 3, in DS Flags bit vector.

```

    0 1 2 3 4 5 6 7
+-----+-----+
| MBZ |G|MBZ|I|N|
+-----+-----+

```


Flag Name and Meaning

G LAG Description Indicator

When this flag is set, DDMAP describes a LAG interface.

7. Interface Index Sub-TLV

The Interface Index object is a Sub-TLV that MAY be included in a DDMAP TLV. Zero or more Interface Index object MAY appear in a DDMAP TLV. The Interface Index Sub-TLV describes the index assigned by the upstream LSR to the interface.

Interface Index Sub-TLV Type is TBD2. Length is 8, and the Value field has following format:

```

      0               1               2               3
      0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
|   Interface Index Flags   |           Must Be Zero           |
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
|                                   Interface Index               |
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+

```

Figure 4: Interface Index Sub-TLV

Interface Index Flags

Interface Index Flags field is a bit vector with following format.

```

      0               1
      0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
| Must Be Zero (Reserved) |M|
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+

```

One flag is defined: M. The remaining flags MUST be set to zero when sending and ignored on receipt.

Flag Name and Meaning

M LAG Member Link Indicator

When this flag is set, interface index described in this sub-TLV is member of a LAG.

Interface Index

Index assigned by the LSR to this interface.

8. Detailed Interface and Label Stack TLV

The Detailed Interface and Label Stack object is a TLV that MAY be included in a MPLS echo reply message to report the interface on which the MPLS echo request message was received and the label stack that was on the packet when it was received. A responder LSR MUST NOT insert more than one instance of this TLV. This TLV allows the initiating LSR to obtain the exact interface and label stack information as it appears at the responder LSR.

Detailed Interface and Label Stack TLV Type is TBD3. Length is K + Sub-TLV Length, and the Value field has following format:

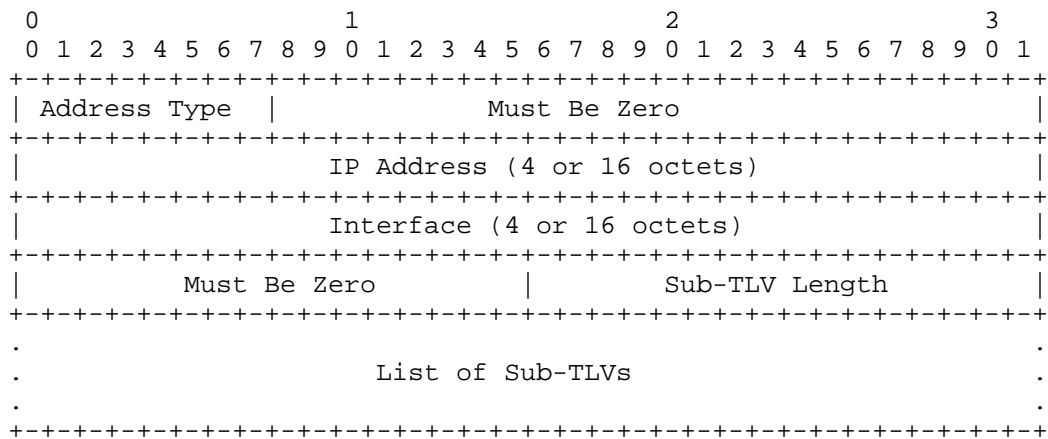


Figure 5: Detailed Interface and Label Stack TLV

The Detailed Interface and Label Stack TLV format is derived from the Interface and Label Stack TLV format (from [RFC4379]). Two changes are introduced. First is that label stack, which is of variable length, is converted into a sub-TLV. Second is that a new sub-TLV is added to describe an interface index. The fields of Detailed Interface and Label Stack TLV have the same use and meaning as in [RFC4379]. A summary of the fields taken from the Interface and Label Stack TLV is as below:

Address Type

The Address Type indicates if the interface is numbered or unnumbered. It also determines the length of the IP Address

and Interface fields. The resulting total for the initial part of the TLV is listed in the table below as "K Octets". The Address Type is set to one of the following values:

Type #	Address Type	K Octets
-----	-----	-----
1	IPv4 Numbered	16
2	IPv4 Unnumbered	16
3	IPv6 Numbered	40
4	IPv6 Unnumbered	28

IP Address and Interface

IPv4 addresses and interface indices are encoded in 4 octets;
IPv6 addresses are encoded in 16 octets.

If the interface upon which the echo request message was received is numbered, then the Address Type MUST be set to IPv4 Numbered or IPv6 Numbered, the IP Address MUST be set to either the LSR's Router ID or the interface address, and the Interface MUST be set to the interface address.

If the interface is unnumbered, the Address Type MUST be either IPv4 Unnumbered or IPv6 Unnumbered, the IP Address MUST be the LSR's Router ID, and the Interface MUST be set to the index assigned to the interface.

Note: Usage of IPv6 Unnumbered has the same issue as [RFC4379], described in Section 3.4.2 of [I-D.ietf-mpls-ipv6-only-gap]. A solution should be considered and applied to both [RFC4379] and this document.

Sub-TLV Length

Total length in octets of the sub-TLVs associated with this TLV.

8.1. Sub-TLVs

This section defines the sub-TLVs that MAY be included as part of the Detailed Interface and Label Stack TLV.

Sub-Type	Value Field
-----	-----
1	Incoming Label stack
2	Incoming Interface Index

8.1.1. Incoming Label Stack Sub-TLV

The Incoming Label Stack sub-TLV contains the label stack as received by the LSR. If any TTL values have been changed by this LSR, they SHOULD be restored.

Incoming Label Stack Sub-TLV Type is 1. Length is variable, and the Value field has following format:

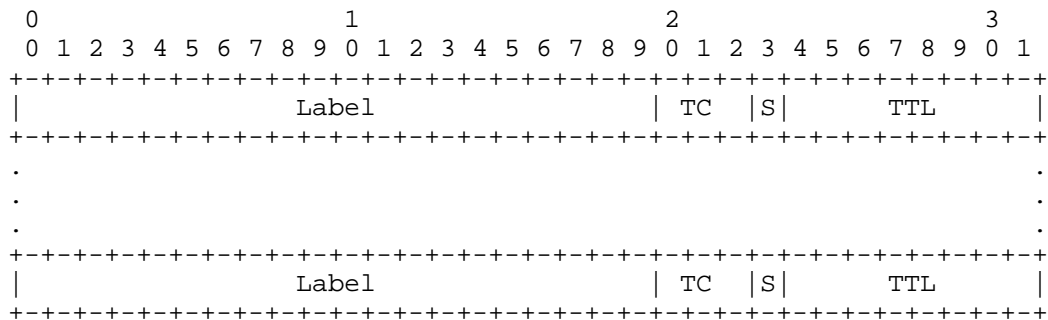


Figure 6: Incoming Label Stack Sub-TLV

8.1.2. Incoming Interface Index Sub-TLV

The Incoming Interface Index object is a Sub-TLV that MAY be included in a Detailed Interface and Label Stack TLV. The Incoming Interface Index Sub-TLV describes the index assigned by this LSR to the interface which received the MPLS echo request message.

Incoming Interface Index Sub-TLV Type is 2. Length is 8, and the Value field has following format:

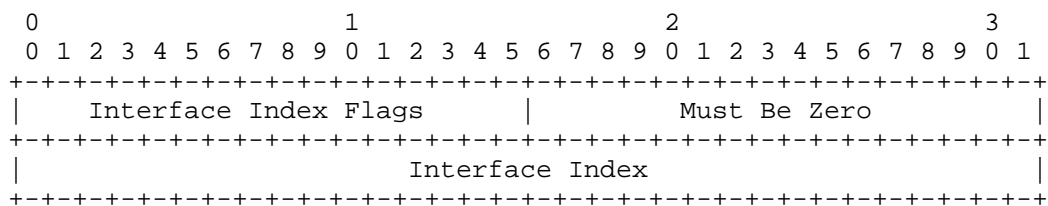


Figure 7: Incoming Interface Index Sub-TLV

Interface Index Flags

Interface Index Flags field is a bit vector with following format.

```

      0                               1
    0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5
+---+---+---+---+---+---+---+---+---+
|   Must Be Zero (Reserved)   |M|
+---+---+---+---+---+---+---+---+

```

One flag is defined: M. The remaining flags MUST be set to zero when sent and ignored on receipt.

Flag	Name and Meaning
M	LAG Member Link Indicator

M LAG Member Link Indicator

When this flag is set, the interface index described in this sub-TLV is a member of a LAG.

Interface Index

Index assigned by the LSR to this interface.

9. Security Considerations

This document extends LSP Traceroute mechanism to discover and exercise layer 2 ECMP paths. Additional processing are required for initiating LSR and responder LSR, especially to compute and handle increasing number of multipath information. Due to additional processing, it is critical that proper security measures described in [RFC4379] and [RFC6424] are followed.

10. IANA Considerations

10.1. LAG Interface Info TLV

The IANA is requested to assign new value TBD1 for LAG Interface Info TLV from the "Multiprotocol Label Switching Architecture (MPLS) Label Switched Paths (LSPs) Ping Parameters - TLVs" registry.

Value	Meaning	Reference
TBD1	LAG Interface Info TLV	this document

10.2. Interface Index Sub-TLV

The IANA is requested to assign new value TBD2 for Interface Index Sub-TLV from the "Multiprotocol Label Switching Architecture (MPLS) Label Switched Paths (LSPs) Ping Parameters - TLVs" registry, "Sub-TLVs for TLV Types 20" sub-registry.

Value	Meaning	Reference
-----	-----	-----
TBD2	Interface Index Sub-TLV	this document

10.3. Detailed Interface and Label Stack TLV

The IANA is requested to assign new value TBD3 for Detailed Interface and Label Stack TLV from the "Multiprotocol Label Switching Architecture (MPLS) Label Switched Paths (LSPs) Ping Parameters - TLVs" registry.

Value	Meaning	Reference
-----	-----	-----
TBD3	Detailed Interface and Label Stack TLV	this document

10.4. New Sub-Registry

10.4.1. DS Flags

[RFC4379] defines the Downstream Mapping TLV, which has the Type 2 assigned from the "Multi-Protocol Label Switching (MPLS) Label Switched Paths (LSPs) Ping Parameters - TLVs" registry. [RFC6424] defines the Downstream Detailed Mapping TLV, which has the Type 20 assigned from the "Multi-Protocol Label Switching (MPLS) Label Switched Paths (LSPs) Ping Parameters - TLVs" registry. DMAP has been deprecated by DDMAP, but both TLVs shares a field: "DS Flags". This document requires allocation of a new value in the "DS Flags" field, which is not maintained by IANA today. Therefore, this document requests IANA to create new registries within [IANA-MPLS-LSP-PING] protocol to maintain "DS Flags" field. Initial values for this registry, "DS Flags", are described below.

Bit number	Name	Reference
-----	-----	-----
7	N: Treat as a Non-IP Packet	RFC4379
6	I: Interface and Label Stack Object Request	RFC4379
5-4	Unassigned	
3	G: LAG Description Indicator	this document
2-0	Unassigned	

Assignments of DS Flags are via Standards Action [RFC5226] or IESG Approval [RFC5226].

Note that "DS Flags" is a field included in two TLVs defined in "Multi-Protocol Label Switching (MPLS) Label Switched Paths (LSPs) Ping Parameters - TLVs" registry: Downstream Mapping TLV (value 2) and Downstream Detailed Mapping TLV (value 20). Modification to "DS Flags" registry will affect both TLVs.

Also note that [I-D.akiya-mpls-entropy-lsp-ping] makes request to create a new retry for "DS Flags", with new values being added for Bit Number 4 and 5. If [I-D.akiya-mpls-entropy-lsp-ping] becomes RFC and "DS Flags" IANA registry is created as result, then this document simply requests Bit Number 3 (G: LAG Description Indicator) to be added to the registry.

10.4.2. Sub-TLVs for TLV Type TBD3

The IANA is requested to make a new "Sub-TLVs for TLV Type TBD3" sub-registry under "Multiprotocol Label Switching Architecture (MPLS) Label Switched Paths (LSPs) Ping Parameters - TLVs" registry. Initial values for this sub-registry, "Sub-TLVs for TLV Types TBD3", are described below.

Sub-Type	Name	Reference
1	Incoming Label Stack	this document
2	Incoming Interface Index	this document
4-65535	Unassigned	

Assignments of Sub-Types are via Standards Action [RFC5226] or IESG Approval [RFC5226].

11. Acknowledgements

TBD

12. References

12.1. 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.
- [RFC6424] Bahadur, N., Kompella, K., and G. Swallow, "Mechanism for Performing Label Switched Path Ping (LSP Ping) over MPLS Tunnels", RFC 6424, November 2011.

12.2. Informative References

[I-D.akiya-mpls-entropy-lsp-ping]

Akiya, N., Swallow, G., and C. Pignataro, "Label Switched Path (LSP) Ping/Trace over MPLS Network using Entropy Labels (EL)", draft-akiya-mpls-entropy-lsp-ping-01 (work in progress), December 2013.

[I-D.ietf-mpls-ipv6-only-gap]

George, W. and C. Pignataro, "Gap Analysis for Operating IPv6-only MPLS Networks", draft-ietf-mpls-ipv6-only-gap-00 (work in progress), April 2014.

[IANA-MPLS-LSP-PING]

IANA, "Multi-Protocol Label Switching (MPLS) Label Switched Paths (LSPs) Ping Parameters",
<<http://www.iana.org/assignments/mpls-lsp-ping-parameters/mpls-lsp-ping-parameters.xhtml>>.

[IEEE802.1AX]

IEEE Std. 802.1AX, "IEEE Standard for Local and metropolitan area networks - Link Aggregation", November 2008.

[RFC5226] Narten, T. and H. Alvestrand, "Guidelines for Writing an IANA Considerations Section in RFCs", BCP 26, RFC 5226, May 2008.

Appendix A. LAG with L2 Switch Issues

Several flavors of "LAG with L2 switch" provisioning models are described in this section, with MPLS data plane ECMP traversal validation issues with each.

A.1. Equal Numbers of LAG Members

R1 ==== S1 ==== R2

The issue with this LAG provisioning model is that packets traversing a LAG member from R1 to S1 can get load balanced by S1 towards R2. Therefore, MPLS echo request messages traversing specific LAG member from R1 to S1 can actually reach R2 via any LAG members, and sender of MPLS echo request messages have no knowledge of this nor no way to control this traversal. In the worst case, MPLS echo request messages with specific entropies to exercise every LAG members from R1 to S1 can all reach R2 via same LAG member. Thus it is impossible for MPLS echo request sender to verify that packets intended to traverse specific LAG member from R1 to S1 did actually traverse that LAG member, and to deterministically exercise "receive" processing of every LAG member on R2.

A.2. Deviating Numbers of LAG Members

R1 ===== S1 ===== R2

There are deviating number of LAG members on the two sides of the L2 switch. The issue with this LAG provisioning model is the same as previous model, sender of MPLS echo request messages have no knowledge of L2 load balance algorithm nor entropy values to control the traversal.

A.3. LAG Only on Right

R1 ---- S1 ===== R2

The issue with this LAG provisioning model is that there is no way for MPLS echo request sender to deterministically exercise both LAG members from S1 to R2. And without such, "receive" processing of R2 on each LAG member cannot be verified.

A.4. LAG Only on Left

R1 ===== S1 ---- R2

MPLS echo request sender has knowledge of how to traverse both LAG members from R1 to S1. However, both types of packets will terminate on the non-LAG interface at R2. It becomes impossible for MPLS echo request sender to know that MPLS echo request messages intended to traverse a specific LAG member from R1 to S1 did indeed traverse that LAG member.

Authors' Addresses

Nobo Akiya
Cisco Systems

Email: nobo@cisco.com

George Swallow
Cisco Systems

Email: swallow@cisco.com

Stephane Litkowski
Orange

Email: stephane.litkowski@orange.com

Bruno Decraene
Orange

Email: bruno.decraene@orange.com

John E. Drake
Juniper Networks

Email: jdrake@juniper.net

MPLS
Internet-Draft
Intended status: Standards Track
Expires: January 4, 2015

S. Bryant
S. Sivabalan
S. Soni
Cisco Systems
July 3, 2014

MPLS Performance Measurement UDP Return Path
draft-bryant-mpls-oam-udp-return-02

Abstract

This document specifies the procedure to be used by the Packet Loss and Delay Measurement for MPLS Networks protocol defined in RFC6374 when sending and processing MPLS performance management out-of-band responses for delay and loss measurements over an IP/UDP return path.

Status of this Memo

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

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

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

This Internet-Draft will expire on January 4, 2015.

Copyright Notice

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

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

1. Introduction

This document describes how Packet Loss and Delay Measurement for MPLS Networks protocol (MPLS-PLDM) [RFC6374] out-of-band responses can be delivered to the Querier using UDP/IP.

The use of UDP may be required to support data path management such as passage through firewalls, or to provide the necessary multiplexing needed in bistatic operation where the querier and the collector are not co-located and the collector is gathering the response information for a number of responders. In a highly scaled system some MPLS-PLDM sessions may be off-loaded to a specific node within a the distributed system that comprises the LSR as a whole. In such systems the response may arrive via any interface in the LSR and need to internally forwarded to the processor tasked with handling the particular MPLS-PLDM measurement. Currently the MPLS-PLDM protocol does not have any mechanism to deliver the PLDM Response message to particular node within a multi-CPU LSR.

The procedure described in this specification describes how the queryer requests delivery of the MPLS-PLDM response over IP to a dynamic UDP port. It makes no other changes to the protocol and thus does not affect the case where the reponse is delivered over a MPLS Associated Channel [RFC5586].

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. Solution Overview

This document specifies that, unless configured otherwise, if a UDP Return Object (URO) is present in a MPLS-PLDM Query, the responder MUST use the IP address and UDP port in the URO to reply back to the querier. Multiple UROs MAY be present in a MPLS-PLDM Query indicating that an identical responses SHOULD be sent to each address-port pair. A responder MAY be designed or configured to only transmit a single response, in which case the response MUST be sent using the parameters specified in the first URO in the query packet.

The procedures defined in this document may be applied to both unidirectional tunnels and bidirectional LSPs. In this document, the term bidirectional LSP includes the co-routed bidirectional LSP defined in [RFC3945] and the associated bidirectional LSP that is

constructed from a pair of unidirectional LSPs (one for each direction) that are associated with one another at the LSP's ingress/egress points [RFC5654]. The mechanisms defined in this document can apply to both IP/MPLS and the MPLS Transport Profile (MPLS-TP).

3.1. UDP Return Object

[Note to reviewers - to be deleted before publication - We considered a number of approaches to the design. The first was to use the existing address object and a separate UDP object, but concern in the WG was that there may be more than one collector that required this information, and the combined size of the two objects. The next approach considered by the authors was to create a new object by appending a UDP port to the existing generalized address object. However by noting that UDP is only likely to be sent over IP and that it will be a long time before we design a third major version of IP we can compress the object either by having separate IPv4 and an IPv4 objects, or using the address length as the discriminator. The object design below uses the latter approach. The resultant combined UDP port + address object is thus the same size as the original address object.]

The format of the UDP Return Object (URO) is as follows:

```

      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
+-----+-----+-----+-----+-----+-----+-----+-----+
| URO TLV Type | Length={6,18} |   UDP-Destination-Port   |
+-----+-----+-----+-----+-----+-----+-----+-----+
~                                         Address                                         ~
+-----+-----+-----+-----+-----+-----+-----+-----+

```

The Type and Length fields are each 8 bits long, and the Length field indicates the size in bytes of the remainder of the object (i.e. is the size of the address in bytes plus 2). When the address is IPv4 the length field is thus 6 and when the address is IPv6 the length field is thus 18. The length field therefore acts as both the TLV parsing parameter and the address family type indicator.

The UDP Return Object Type (URO TLV Type) has a value of <TBD>.

The UDP Destination Port is a UDP Destination port as specified in [RFC0768].

The Address is either an IPv4 or an IPv6 address.

The URO MUST NOT appear in a response.

4. Theory of Operation

This document defines the UDP Return Object to enable the MPLS-PLDM Querier to specify the return path for the MPLS-PLDM reply using IP/UDP encapsulation.

When the MPLS-PLDM Response is requested out-of-band by setting Control Code of the MPLS-PLDM Query to "Out-of-band Response Requested", and the URO is present, the responder SHOULD send the response back to Querier on the specified destination UDP port at the specified destination IP address as received in the URO.

If the URO is expected but is not present in Query message and an MPLS-PLDM Response is requested out-of-band, the Query message MUST NOT be processed further. If received over a bidirectional LSP, the control code of the Response message MUST be set to "Error - Missing TLV" and a Response SHOULD be sent over the reverse LSP. The receipt of such a mal-formed request SHOULD be notified to the operator through the management system, taking the normal precautions with respect to the prevention of overload of the error reporting system.

4.1. Missing TLV

The control code "Missing TLV", which is classified as an Error response code, indicates that the operation failed because one or more required TLV Objects was not sent in the query message.

4.2. Sending an MPLS-PM Query

When sending an MPLS-PLDM Query message, in addition to the rules and procedures defined in [RFC6374]; the Control Code of the MPLS-PLDM Query MUST be set to "Out-of-band Response Requested", and a URO MUST be carried in the MPLS-PLDM Query message.

If the Querier uses the UDP port to de-multiplexing of the response for different measurement type, there MUST be a different UDP port for each measurement type (Delay, loss and delay-loss combined).

An implementation MAY use multiple UDP ports for same measurement type to direct the response to the correct management process in the LSR.

4.3. Receiving an MPLS PM Query Request

The processing of MPLS-PLDM query messages as defined in [RFC6374] applies in this document. In addition, when an MPLS-PLDM Query request is received, with the Control Code of the MPLS-PLDM Query set to "Out-of-band Response Requested" with a URO present, then the

Responder SHOULD use that IP address and UDP port to send MPLS-PLDM response back to Querier.

If an Out-of-band response is requested and the Address object or the URO is missing, the Query SHOULD be dropped in the case of a unidirectional LSP. If both these TLVs are missing on a bidirectional LSP, the control code of Response message should set to "Invalid Message" and the response SHOULD be sent over the reverse LSP. The receipt of such a mal-formed request SHOULD be notified to the operator through the management system, taking the normal precautions with respect to the prevention of overload of the error reporting system.

4.4. Sending an MPLS-PM Response

As specified in [RFC6374] the MPLS-PLDM Response can be sent over either the reverse MPLS LSP for a bidirectional LSP or over an IP path. It MUST NOT be sent other than in response to an MPLS-PLDM Query message.

When the requested return path is an IP forwarding path and this method is in use, the destination IP address and UDP port MUST be copied from the URO. The source IP address and the source UDP port of Response packet is left to discretion of the Responder subject to the normal management and security considerations. The packet format for the MPLS-PLDM response after the UDP header is as specified in [RFC6374]. As shown in Figure 1 the Associate Channel Header (ACH) [RFC5586] is not included. The information provided by the ACH is not needed since the correct binding between the Query and Response messages is achieved through the UDP Port and the Session Identifier contained in the RFC6374 message.

```
+-----+
|  IP Header                               |
|  .   Source Address = Responders IP Address   |
|  .   Destination Address = URO.Address         |
|  .   Protocol = UDP                           |
|  .                                             |
+-----+
|  UDP Header                               |
|  .   Source Port = As chosen by Responder      |
|  .   Destination Port = URO.UDP-Destination-Port |
|  .                                             |
+-----+
|  Message as specified in RFC6374           |
|  .                                             |
+-----+
```

Figure 1: Response packet Format

If the return path is an IP path, only one-way delay or one-way loss measurement can be carried out. In this case timestamps 3 and 4 MUST be zero as specified in [RFC6374].

4.5. Receiving an MPLS-PM Response

If the response was received over UDP/IP and an out-of-band response was expected, the Response message SHOULD be directed to the appropriate measurement process as determined by the destination UDP Port, and processed using the corresponding measurement type procedure specified in [RFC6374].

If the Response was received over UDP/IP and an out-of-band response was not requested, that response should be dropped and the event SHOULD be notified to the operator through the management system, taking the normal precautions with respect to the prevention of overload of the error reporting system.

5. Manageability Considerations

The manageability considerations described in Section 7 of [RFC6374] are applicable to this specification. Additional manageability considerations are noted within the elements of procedure of this document.

Nothing in this document precludes the use of a configured UDP/IP return path in a deployment in which configuration is preferred to signalling. In these circumstances the URO MAY be omitted from the

MPLS-PLDM messages.

6. Security Considerations

The MPLS-PLDM system is not intended to be deployed on the public Internet. It is intended for deployment in well managed private and service provider networks. The security considerations described in Section 8 of [RFC6374] are applicable to this specification and the reader's attention is drawn to the last two paragraphs. Cryptographic measures may be enhanced by the correct configuration of access control lists and firewalls.

There is no additional exposure of information to pervasive monitoring systems observing LSPs that are being monitored.

7. IANA Considerations

IANA is requested to assign a new Optional TLV type from MPLS Loss/Delay Measurement TLV Object Registry contained within the g-ach-parameters registry set.

Code	Description	Reference
TBD	Return UDP Port	[This]

The TLV 131 is recommended.

IANA is requested to assign a new response code in the MPLS Loss/Delay Measurement Control Code Registry contained within the g-ach-parameters registry set.

Code	Description	Reference
TBD	Missing TLV	[This]

The response code 0x1E is recommended.

8. Acknowledgements

We acknowledge the contribution of Joseph Chin and Rakesh Gandhi, both with Cisco Systems. We thank Loa Andersson, Eric Osborne, Mustapha Aissaoui, and Jeffrey Zhang for their review comments.

We thank all who have reviewed this text and provided feedback.

9. Normative References

[RFC0768] Postel, J., "User Datagram Protocol", STD 6, RFC 768,

August 1980.

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC3945] Mannie, E., "Generalized Multi-Protocol Label Switching (GMPLS) Architecture", RFC 3945, October 2004.
- [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.
- [RFC6374] Frost, D. and S. Bryant, "Packet Loss and Delay Measurement for MPLS Networks", RFC 6374, September 2011.

Authors' Addresses

Stewart Bryant
Cisco Systems

Email: stbryant@cisco.com

Siva Sivabalan
Cisco Systems

Email: msiva@cisco.com

Sagar Soni
Cisco Systems

Email: sagsoni@cisco.com

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: January 3, 2015

WQ. Cheng
L. Wang
H. Li
China Mobile
H. van Helvoort
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 02, 2014

MPLS-TP Shared-Ring protection (MSRP) mechanism for ring topology
draft-cheng-mpls-tp-shared-ring-protection-02

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 RFC6564 [RFC5654] for optimized ring protection.

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 [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 3, 2015.

Copyright Notice

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

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

Table of Contents

1. Introduction	3
1.1. Requirements for MPLS-TP ring protection	4
1.1.1. Recovery for Multiple failures	4
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	8
2.1.3. Failure detection	9
2.2. P2P wrapping	9
2.2.1. Wrapping for Link Failure	10
2.2.2. Wrapping for node Failure	10
2.3. P2P short wrapping	11
2.4. P2P steering	12
2.5. P2P wrapping for Interconnected Rings	15
2.5.1. Interconnected ring topology	15
2.5.2. Interconnected ring protection scheme	16
3. Coordination protocol	21
3.1. RPS protocol	21
3.1.1. Transmission and acceptance of RPS requests	23
3.1.2. RPS PDU structure	23
3.1.3. Ring node RPS states	24
3.1.4. RPS state transitions	26
3.2. APS state machine	28
3.2.1. Initial states	28

3.2.2.	State transitions when local request is applied . . .	29
3.2.3.	State transitions when remote request is applied . .	33
3.2.4.	State Transitions when request addresses to another node is received	35
4.	IANA Considerations	38
5.	Security Considerations	39
6.	Normative Refereces	39
	Authors' Addresses	39

1. Introduction

As described in 2.5.6.1. ring protection of MPLS-TP requirements RFC5654 [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 [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 [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 [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 [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 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, Minxue Wang, Sheng Liu (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.

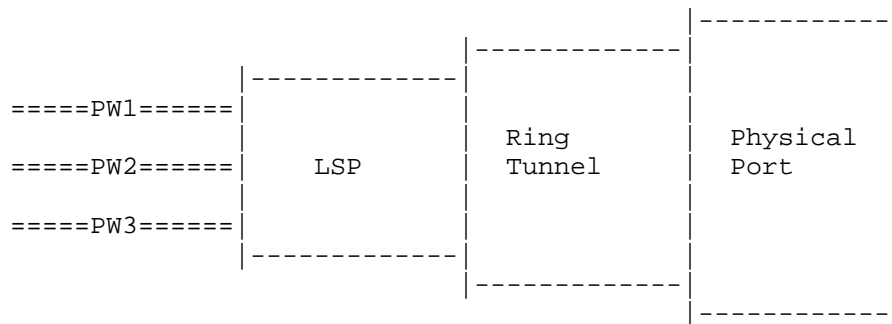


Figure 1 the logic layers of the ring

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

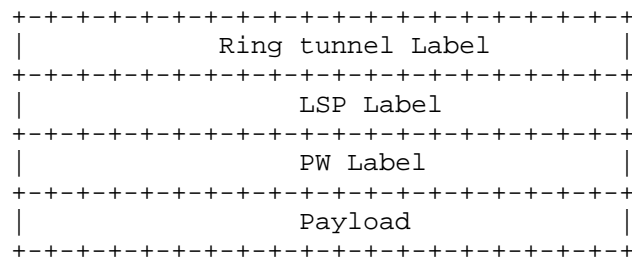


Figure 2 Label stack used in MPLS-TP Shared Ring Protection

2.1.1.1. 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;

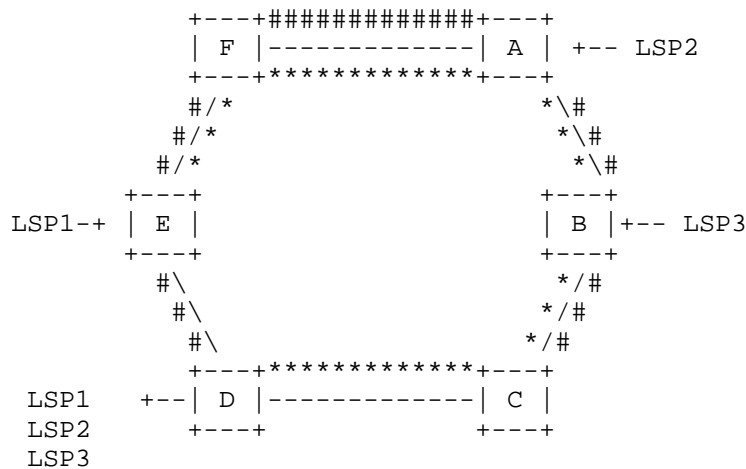
To Node B: RcW_B, RaW_B, RcP_B, RaP_B;

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.



---- physical links

**** RcW_D

RaP_D

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 RFC5654 [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.

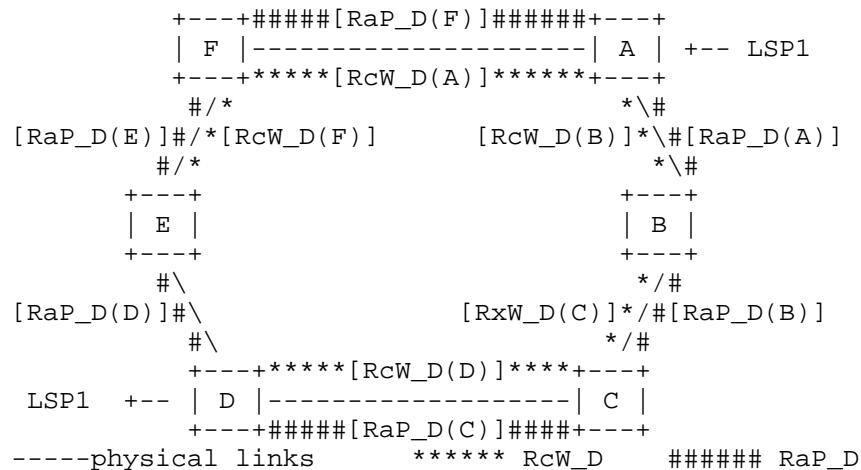


Figure 4 Label operation of MSRP

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 [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).

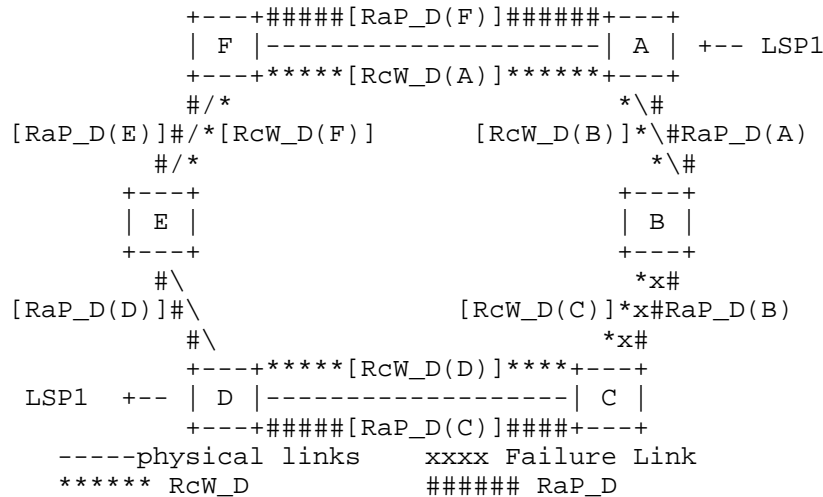


Figure 5 P2P wrapping for link failure in a single ring

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).

```

+---+#####[RaP_D(F)]#####+---+
| F |-----| A | +--- LSP1
+---+*****[RcW_D(A)]*****+---+
#/*                               *\#
[RaP_D(E)]#/*[RcW_D(F)]          [RcW_D(B)]*\#RaP_D(A)
#/*                               *\#
+---+                               xxxxx
| E |                               x B x
+---+                               xxxxx
#\                               */#
[RaP_D(D)]#\                      [RcW_D(C)]*/#RaP_D(B)
#\                               */#
+---+*****[RcW_D(D)]*****+---+
LSP1  +--- | D |-----| C |
+---+#####[RaP_D(C)]#####+---+
-----physical links          xxxxxx Failure Node
*****RcW_D                    ##### RaP_D

```

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.

```

+---+#####[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#
+---+*****[RcW_D(D)]*****+---+
LSP1 +-- | D |-----| C |
+---+                                     +---+
-----physical links      xxxx Failure Link
***** RcW_D              ##### RaP_D

```

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.

```

+---LSP 1
+---+ +---+ ##### [RaP_D(F)] ##### +---+ +---+
|F|A|B|C|D|E|F| | F | ----- | A | |A|B|C|D|E|F|A|
+---+ +---+ ***[RcW_D(A)]*** +---+ +---+
|I|I|I|S|I|I| |I|I|S|I|I|I|
+---+ +---+ #/* * \# +---+
[RaP_D(E)] #/* [RcW_D(B)] * \# [RaP_D(A)]
#/* [RcW_D(F)] * \#
+---+ +---+ #/* * \#
|E|F|A|B|C|D|E| +---+ +---+ LSP 2
+---+ +---+ | E | +---+ | B | +---+
|I|I|I|I|S|I| +---+ |B|C|D|E|F|A|B|
+---+ +---+ # \# * /# +---+
# \# [RcW_D(E)] [RcW_D(C)] * /# |I|S|I|I|I|I|
[RaP_D(D)] # \# * /# +---+
# \# * /# [RaP_D(B)]
+---+ +---+ [RcW_D(D)] +---+ +---+
|D|E|F|A|B|C|D| +---+ | D | xxxxxxxxxxxxxxxxxxxx | C | |C|D|E|F|A|B|C|
+---+ +---+ LSP 1 +---+ [RaP_D(C)] +---+ +---+
|I|I|I|I|I|S| LSP 2 |S|I|I|I|I|I|
+---+ +---+

```

----- physical links ***** RcW_D ##### RaP_D

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) .

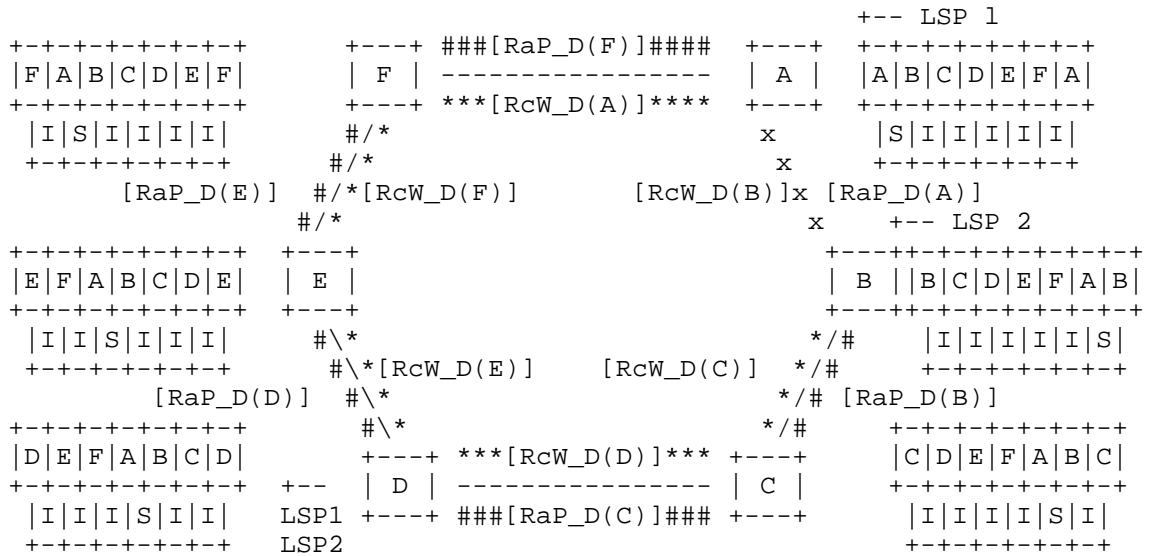
If the link between C and D breaks down, as Figure 8 shows, according to the fault detection function of each link, Node D will find out that there is a failure in the link between C and D, and it will update the link state of its ring topology, changing the link state between C and D from normal to 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.



----- physical links ***** RcW_D ##### RaP_D
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.

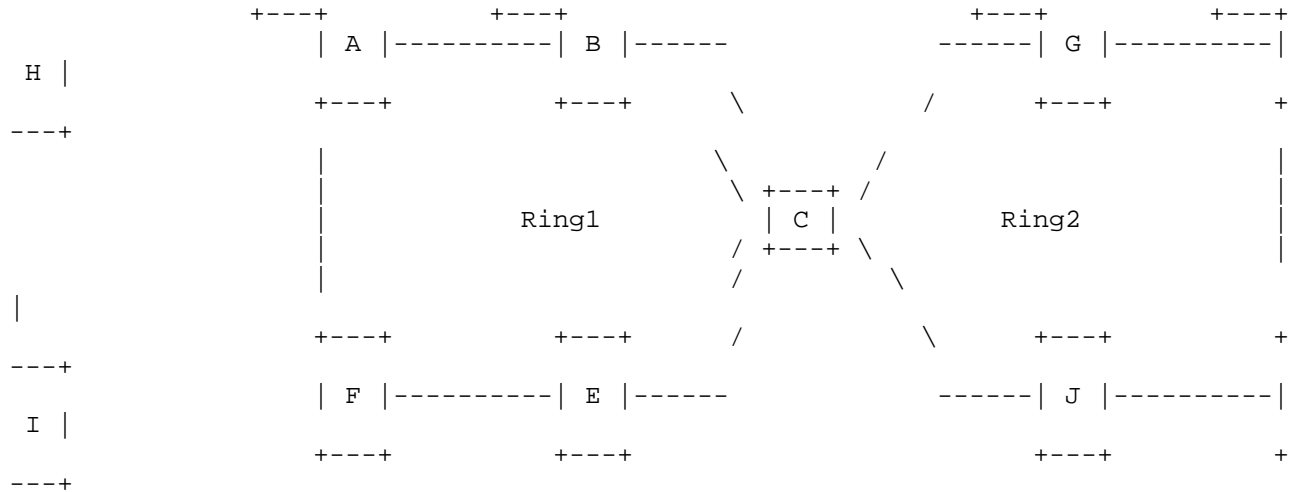


Figure 10 Single-node interconnected rings

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.

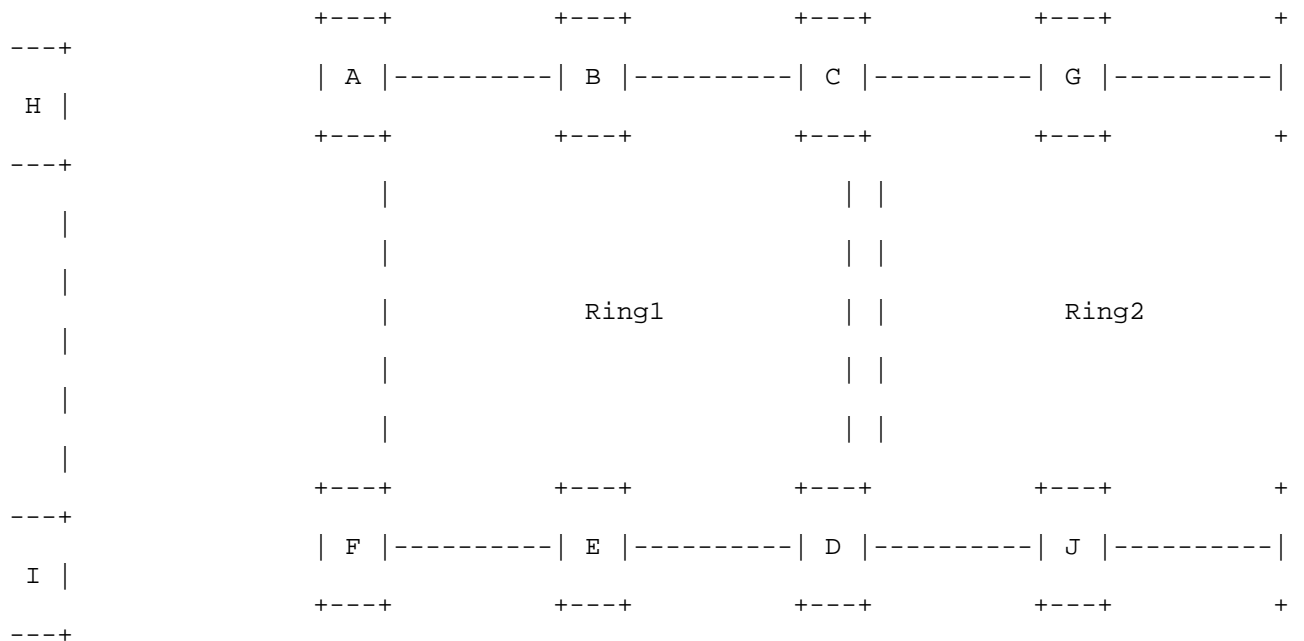


Figure 11 Dual-node interconnected rings

2.5.2. Interconnected ring protection scheme

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

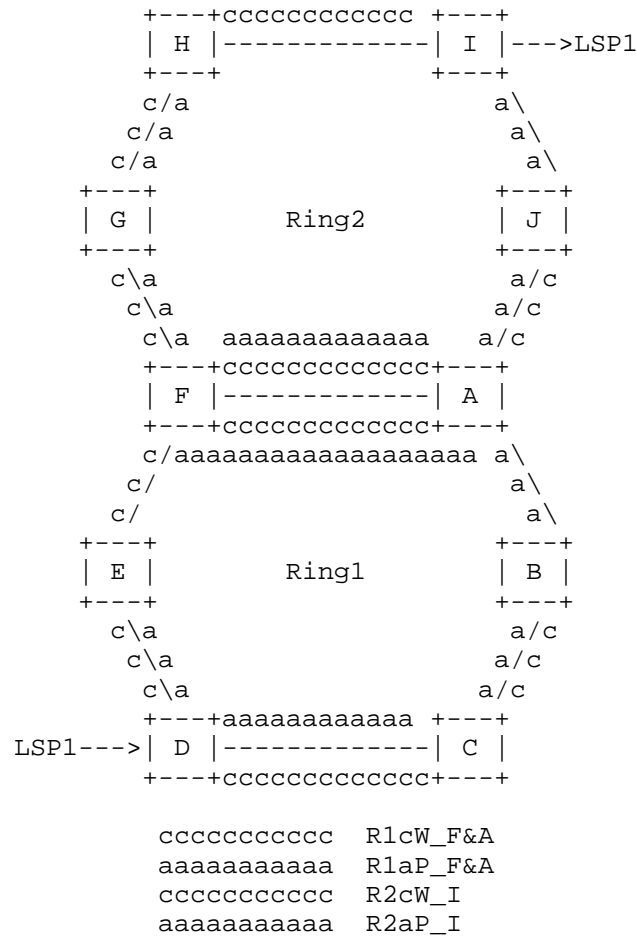


Figure 12 Ring tunnels for the interconnected rings

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->E)->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.

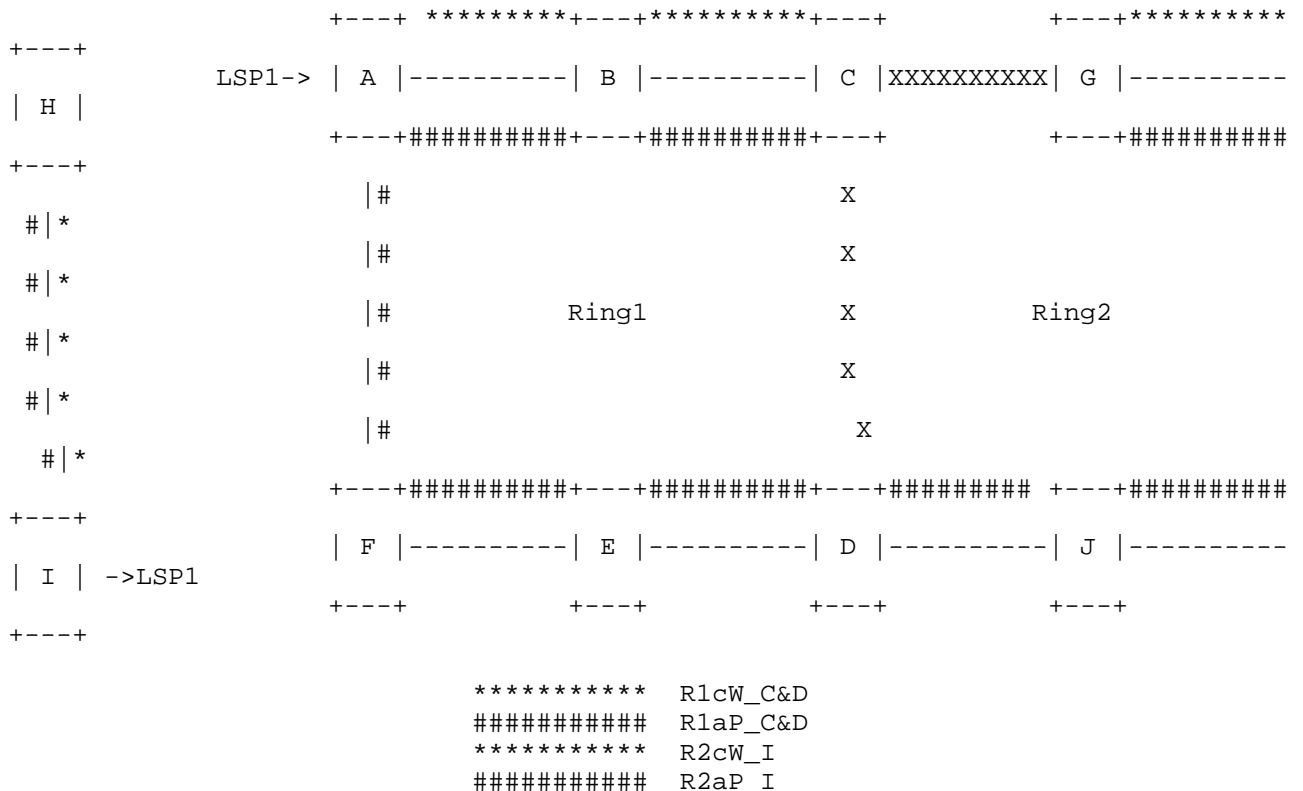


Figure 13 interconnected ring

3. Coordination protocol

3.1. RPS protocol

The MSRP protection operation MUST be controlled with the help of the Ring Protection Switch Protocol(RPS). The RPS processes in the each of the individual nodes that form the ring SHOULD communicate using G-ACh channel.

The RPS protocol MUST carry the ring status information and RPS requests, both automatic and externally initiated commands, between the ring nodes.

Each node on the ring MUST be uniquely identified by assigning it a node ID. The maximum number of nodes on the ring supported by the RPS protocol is 127. The node ID SHOULD be independent of the order in which the nodes appear on the ring. The node ID is used to identify the source and destination nodes of each RPS request.

Each node SHOULD have a ring map containing information about the sequence of the nodes around the ring. The method of configuring the nodes with the ring maps is TBD.

When no protection switches are active on the ring, each node MUST dispatch periodically RPS requests to the two adjacent nodes, indicating No Request (NR). When a node determines that a protection switching is required, it MUST send the appropriate RPS request in both directions.

```

          +----+ A->B(NR)      +----+ B->C(NR)      +----+ C->D(NR)
-----|  A  |-----|  B  |-----|  C  |-----
(NR)F<-A +----+      (NR)A<-B +----+      (NR)B<-C +----+

```

Figure 14: RPS communication between the ring nodes in case of no failures in the ring

A destination node is a node that is adjacent to a node that identified a failed span. When a node that is not the destination node receives an RPS request and it has no higher priority local request, it MUST transfer the RPS request as received. In this way, the switching nodes can maintain direct RPS protocol communication in the ring.

```

          +----+ C->B(SF)      +----+ B->C(SF)      +----+ C->B(SF)
-----|  A  |-----|  B  |-----|  C  |-----
(SF)C<-B +----+      (SF)C<-B +----+      (SF)B<-C +----+

```

Figure 15: RPS communication between the ring nodes in case of failure between nodes B and C

Note that in the case of a bidirectional failure such as a cable cut, two nodes detect the failure and send each other an RPS request in opposite directions.

o In rings utilizing the wrapping protection, when the destination node receives the RPS request it MUST perform the switch from/to the working ring tunnels to/from the protection ring tunnels if it has no higher priority active APS request.

o In rings utilizing the steering protection, when a ring switch is required, any node MUST perform the switches if its added/dropped traffic is affected by the failure. Determination of the affected traffic SHOULD be performed by examining the APS requests (indicating the nodes adjacent to the failure or failures) and the stored ring maps (indicating the relative position of the failure and the added traffic destined towards that failure).

When the failure has cleared and the Wait-to-Restore (WTR) timer has expired, the nodes sourcing RPS requests MUST drop their respective

switches (tail end) and MUST source an RPS request carrying NR code. The node receiving from both directions such RPS request (head end) MUST drop its protection switches.

A protection switch MUST be initiated by one of the criteria specified in Section 3.2. A failure of the RPS protocol or controller MUST NOT trigger a protection switch.

Ring switches MUST be preempted by higher priority RPS requests. For example, consider a protection switch that is active due to a manual switch request on the given span, and another protection switch is required due to a failure on another span. Then a RPS request MUST be generated, the former protection switch MUST be dropped, and the latter protection switch established.

MSPP mechanism SHOULD support multiple protection switches in the ring, resulting the ring being segmented into two or more separate segments. This may happen when several APS requests of the same priority exist in the ring due to multiple failures or external switch commands.

Proper operation of the MSRP mechanism relies on all nodes having knowledge of the state of the ring (nodes and spans) so that nodes do not preempt existing RPS request unless they have a higher-priority RPS request. In order to accommodate ring state knowledge, during protection switch the RPS requests MUST be sent in both directions.

3.1.1. Transmission and acceptance of RPS requests

A new RPS request MUST be transmitted immediately when a change in the transmitted status occurs.

The first three RPS protocol messages carrying new RPS request SHOULD be transmitted as fast as possible. For fast protection switching within 50 ms, the interval of the first three RPS protocol messages SHOULD be 3.3 ms. Then RPS requests SHOULD be transmitted with the interval of 5 seconds.

3.1.2. RPS PDU structure

Figure 16 depicts the format of an RPS packet that is sent on the G-ACh. The Channel Type field is set to indicate that the message is an RPS message. The ACH MUST NOT include the ACH TLV Header RFC5586 [RFC5586] meaning that no ACH TLVs can be included in the message.

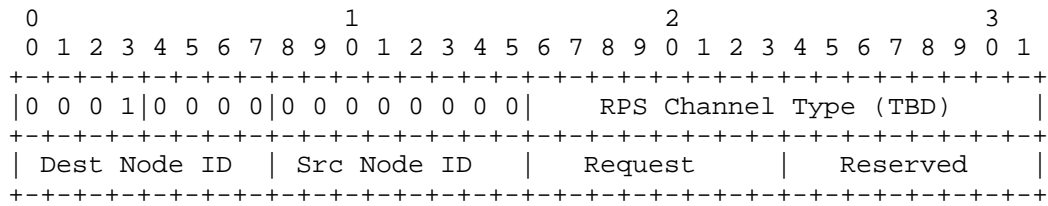


Figure 16: G-ACh RPS Packet

The following fields MUST be provided:

o Destination Node ID: The destination node ID MUST always be set to value of a node ID of the adjacent node. Valid destination node ID values are 1-127.

o Source node ID: The source node ID MUST always be set to the value of the node ID generating the APS request. Valid source node ID values are 1-127.

o RPS request code: A code consisting of four bits as specified below.

Bits 4-1 (MSB - LSB)	Condition, State or external Request	Priority
1 1 1 1	Lockout of Protection (LP)	highest
1 1 0 1	Forced Switch (FS)	
1 0 1 1	Signal Fail (SF)	
0 1 1 0	Manual Switch (MS)	
0 1 0 1	Wait-To-Restore (WTR)	
0 0 1 1	Exerciser (EXER)	
0 0 0 1	Reverse Request (RR)	
0 0 0 0	No Request (NR)	lowest

3.1.3. Ring node RPS states

Idle state: A node is in the idle state when it has no RPS request and is sourcing and receiving NR code to/from both directions.

Switching state: A node not in the idle or pass-through states is in the switching state.

Pass-through state: A node is in the pass-through state when its highest priority RPS request is a request not destined to or sourced by it. The pass-through is bidirectional.

3.1.3.1. Idle state

A node in the idle state MUST source the NR request in both directions.

A node in the idle state MUST terminate RPS requests flow in both directions.

A node in the idle state MUST block the traffic flow on protection LSPs/tunnels in both directions.

3.1.3.2. Switching state

A node in the switching state MUST source RPS request to adjacent node with its highest RPS request code in both directions when it detects a failure or receives an external command.

A node in the switching state MUST terminate RPS requests flow in both directions.

As soon as it receives an RPS request from the short path, the node to which it is addressed MUST acknowledge the RPS request by replying with the RR code on the short path, and with the received RPS request code on the long path.

This rule refers to the unidirectional failure detection: the RR SHOULD be issued only when the node does not detect the failure condition (i.e., the node is a head end), that is, it is not applicable when a failure is detected bidirectionally, because, in this latter case, both nodes send an RPS request for the failure on both paths (short and long).

The following switches MUST be allowed to coexist:

- o LP with LP
- o FS with FS
- o SF with SF
- o FS with SF

When multiple MS RPS requests over different spans exist at the same time, no switch SHOULD be executed and existing switches MUST be dropped. The nodes MUST signal, anyway, the MS APS request code.

Multiple EXER request MUST be allowed to coexist in the ring.

A node in a ring switching state that receives the external command LW for the affected span MUST drop its switch and MUST signal NR for the locked span if there is no other APS request on another span. Node still SHOULD signal relevant APS request for another span.

3.1.3.3. Pass-through state

When a node is in a pass-through state, it MUST transmit on one side, the same RPS request as it receives from the other side.

When a node is in a pass-through state, it MUST allow the traffic flow on protection ring tunnels in both directions.

3.1.4. RPS state transitions

All state transitions are triggered by an incoming APS request change, a WTR expiration, an externally initiated command, or locally detected MPLS-TP section failure conditions.

RPS requests due to a locally detected failure, an externally initiated command, or received APS request shall pre-empt existing RPS requests in the prioritized order given in Section 3.1.2, unless the requests are allowed to coexist.

3.1.4.1. Transitions between the idle and pass-through states

The transition from the idle state to pass-through state MUST be triggered by a valid APS request change, in any direction, from the NR code to any other code, as long as the new request is not destined for the node itself. Both directions move then into a pass-through state, so that, traffic entering the node through the protection Ring tunnels are by-passed across the node.

A node MUST revert from pass-through state to the idle state when it detects NR codes incoming from both directions. Both directions revert simultaneously from the pass-through state to the idle state.

3.1.4.2. Transitions between the idle and switching states

Transition of a node from the idle state to the switching state MUST be triggered by one of the following conditions:

- o a valid RPS request change from the NR code to any code received on either the long or the short path and destined to this node
- o an externally initiated command for this node

o the detection of an MPLS-TP section layer failure at this node. Actions taken at a node in idle state upon transition to switching state are:

o for all protection switch requests, except EXER and LP, the node MUST execute the switch

o for EXER, and LP, the node MUST signal appropriate request but not execute the switch.

A node MUST revert from the switching state to the idle state when it detects NR codes received from both directions.

o At the tail end: When a WTR time expires or an externally initiated command is cleared at a node, the node MUST drop its switch, transit to Idle state and signal the NR code in both directions.

o At the head end: Upon reception of the NR code, from both directions, the head-end node MUST drop its switch, transition to Idle state and signal the NR code in both directions.

3.1.4.3. Transitions between switching states

When a node that is currently executing any protection switch receives a higher priority APS request (due to a locally detected failure, an externally initiated command, or a ring protection switch request destined to it) for the same span, it MUST upgrade the priority of the switch it is executing to the priority of the received APS request.

When a failure condition clears at a node, the node MUST enter WTR condition and remain in it for the appropriate time-out interval, unless:

o a different RPS request of higher priority than WTR is received

o another failure is detected

o an externally initiated command becomes active.

The node MUST send out a WTR code on both the long and short paths.

When a node that is executing a switch in response to incoming SF RPS request (not due to a locally detected failure) receives a WTR code (unidirectional failure case), it MUST send out RR code on the short path and the WTR on the long path.

3.1.4.4. Transitions between switching and pass-through states

When a node that is currently executing a switch receives an RPS request for a non-adjacent span of higher priority than the switch it is executing, it **MUST** drop its switch immediately and enter the pass-through state.

The transition of a node from pass-through to switching state **MUST** be triggered by:

- o an equal, higher priority, or allowed coexisting externally initiated command
- o the detection of an equal, higher priority, or allowed coexisting failure
- o the receipt of an equal, higher priority, or allowed coexisting APS request destined to this node.

3.2. APS state machine

3.2.1. Initial states

State		Signaled APS
A	Idle Working: no switch Protection: no switch	NR
B	Pass-trough Working: no switch Protection: pass through	N/A
C	Switching - LP Working: no switch Protection: no switch	LP
D	Idle - LW Working: no switch Protection: no switch	NR
E	Switching - FS Working: switched Protection: switched	FS
F	Switching - SF Working: switched Protection: switched	SF
G	Switching - MS Working: switched Protection: switched	MS
H	Switching - WTR Working: switched Protection: switched	WTR
I	Switching - EXER Working: no switch Protection: no switch	EXER

3.2.2. State transitions when local request is applied

In the state description below 'O' means that new local request will be rejected because of exiting request.

```
=====
Initial state      New request      New state
-----
```

A (Idle)	LP	C (Switching - LP)
	LW	D (Idle - LW)
	FS	E (Switching - FS)
	SF	F (Switching - SF)
	Recover from SF	N/A
	MS	G (Switching - MS)
	Clear	N/A
	WTR expires	N/A
	EXER	I (Switching - EXER)
Initial state	New request	New state
-----	-----	-----
B (Pass-trough)	LP	C (Switching - LP)
	LW	B (Pass-trough)
	FS	O - if current state is due to LP sent by another node
		E (Switching - FS) - otherwise
	SF	O - if current state is due to LP sent by another node
		F (Switching - SF) - otherwise
	Recover from SF	N/A
	MS	O - if current state is due to LP, SF or FS sent by another node
		G (Switching - MS) - otherwise
	Clear	N/A
	WTR expires	N/A
	EXER	O
Initial state	New request	New state
-----	-----	-----
C (Switching - LP)	LP	N/A
	LW	O
	FS	O
	SF	O
	Recover from SF	N/A
	MS	O
	Clear	A (Idle) - if there is no failure in the ring
		F (Switching - SF) - if there is a failure at this node
		B (Pass-trough) - if there is a failure at another node
	WTR expires	N/A
	EXER	O
Initial state	New request	New state
-----	-----	-----

D (Idle - LW)	LP	C (Switching - LP)
	LW	N/A - if on the same span
		D (Idle - LW) - if on another span
	FS	O - if on the same span
		E (Switching - FS) - if on another span
	SF	O - if on the addressed span
		F (Switching - SF) - if on another span
	Recover from SF	N/A
	MS	O - if on the same span
		G (Switching - MS) - if on another span
	Clear	A (Idle) - if there is no failure on addressed span
		F (Switching - SF) - if there is a failure on this span
	WTR expires	N/A
	EXER	O
=====		
Initial state	New request	New state
-----	-----	-----
E (Switching - FS)	LP	C (Switching - LP)
	LW	O - if on another span
		D (Idle - LW) - if on the same span
	FS	N/A - if on the same span
		E (Switching - FS) - if on another span
	SF	O - if on the addressed span
		E (Switching - FS) - if on another span
	Recover from SF	N/A
	MS	O
	Clear	A (Idle) - if there is no failure in the ring
		F (Switching - SF) - if there is a failure at this node
		B (Pass-through) - if there is a failure at another node
	WTR expires	N/A
	EXER	O
=====		
Initial state	New request	New state
-----	-----	-----
F (Switching - SF)	LP	C (Switching - LP)
	LW	O - if on another span

		D (Idle - LW) - if on the same span
	FS	E (Switching - FS)
	SF	N/A - if on the same span
		F (Switching - SF) - if on another span
	Recover from SF	H (Switching - WTR)
	MS	O
	Clear	N/A
	WTR expires	N/A
	EXER	O
=====		
Initial state	New request	New state
-----	-----	-----
G (Switching - MS)	LP	C (Switching - LP)
	LW	O - if on another span
		D (Idle - LW) - if on the same span
	FS	E (Switching - FS)
	SF	F (Switching - SF)
	Recover from SF	N/A
	MS	N/A - if on the same span
		G (Switching - MS) - if on another span release the switches but signal MS
	Clear	A
	WTR expires	N/A
	EXER	O
=====		
Initial state	New request	New state
-----	-----	-----
H (Switching - WTR)	LP	C (Switching - LP)
	LW	D (Idle - W)
	FS	E (Switching - FS)
	SF	F (Switching - SF)
	Recover from SF	N/A
	MS	G (Switching - MS)
	Clear	A
	WTR expires	A
	EXER	O
=====		
Initial state	New request	New state
-----	-----	-----
I (Switching - EXER)	LP	C (Switching - LP)
	LW	D (idle - W)
	FS	E (Switching - FS)
	SF	F (Switching - SF)
	Recover from SF	N/A

MS	G (Switching - MS)
Clear	A
WTR expires	N/A
EXER	N/A - if on the same span
	I (Switching - EXER)

=====

3.2.3. State transitions when remote request is applied

The priority of remote request does not depend on the side from which the request is received.

Initial state	New request	New state
A (Idle)	LP	C (Switching - LP)
	FS	E (Switching - FS)
	SF	F (Switching - SF)
	MS	G (Switching - MS)
	WTR	N/A
	EXER	I (Switching - EXER)
	RR	N/A
	NR	A (Idle)

=====

Initial state	New request	New state
B (Pass-through)	LP	C (Switching - LP)
	FS	N/A - cannot happen when there is LP request in the ring
		E (Switching - FS) - otherwise
	SF	N/A - cannot happen when there is LP request in the ring
		F (Switching - SF) - otherwise
	MS	N/A - cannot happen when there is LP, FS or SF request in the ring
		G (Switching - MS) - otherwise
	WTR	N/A - cannot happen when there is LP, FS, SF or MS request in the ring
	EXER	N/A - cannot happen when there is LP, FS, SF, MS or WTR request in the ring
		I (Switching - EXER) - otherwise
	RR	N/A
	NR	A (Idle) - if received from both sides

=====		
Initial state	New request	New state
-----	-----	-----
C (Switching - LP)	LP	C (Switching - LP)
	FS	N/A - cannot happen when there is LP request in the ring
	SF	N/A - cannot happen when there is LP request in the ring
	MS	N/A - cannot happen when there is LP request in the ring
	WTR	N/A
	EXER	N/A - cannot happen when there is LP request in the ring
	RR	C (Switching - LP)
	NR	N/A
=====		
Initial state	New request	New state
-----	-----	-----
D (Idle - LW)	LP	C (Switching - LP)
	FS	E (Switching - FS)
	SF	F (Switching - SF)
	MS	G (Switching - MS)
	WTR	N/A
	EXER	I (Switching - EXER)
	RR	N/A
	NR	D (Idle - LW)
=====		
Initial state	New request	New state
-----	-----	-----
E (Switching - FS)	LP	C (Switching - LP)
	FS	E (Switching - FS)
	SF	E (Switching - FS)
	MS	N/A - cannot happen when there is FS request in the ring
	WTR	N/A
	EXER	N/A - cannot happen when there is FS request in the ring
	RR	E (Switching - FS)
	NR	N/A
=====		
Initial state	New request	New state
-----	-----	-----
F (Switching - SF)	LP	C (Switching - LP)
	FS	F (Switching - SF)
	SF	F (Switching - SF)
	MS	N/A - cannot happen when there is SF request in the ring
	WTR	N/A

	EXER	N/A - cannot happen when there is SF request in the ring
	RR	F (Switching - SF)
	NR	N/A
=====		
Initial state	New request	New state
-----	-----	-----
G (Switching - MS)	LP	C (Switching - LP)
	FS	E (Switching - FS)
	SF	F (Switching - SF)
	MS	G (Switching - MS) - release the switches but signal MS
	WTR	N/A
	EXER	N/A - cannot happen when there is MS request in the ring
	RR	G (Switching - MS)
	NR	N/A
=====		
Initial state	New request	New state
-----	-----	-----
H (Switching - WTR)	LP	C (Switching - LP)
	FS	E (Switching - FS)
	SF	F (Switching - SF)
	MS	G (Switching - MS)
	WTR	H (Switching - WTR)
	EXER	N/A - cannot happen when there is WTR request in the ring
	RR	H (Switching - WTR)
	NR	N/A
=====		
Initial state	New request	New state
-----	-----	-----
I (Switching - EXER)	LP	C (Switching - LP)
	FS	E (Switching - FS)
	SF	F (Switching - SF)
	MS	G (Switching - MS)
	WTR	N/A
	EXER	I (Switching - EXER)
	RR	I (Switching - EXER)
	NR	N/A
=====		

3.2.4. State Transitions when request addresses to another node is received

The priority of remote request does not depend on the side from which the request is received.

Initial state	New request	New state
A (Idle)	LP	B (Pass-trough)
	FS	B (Pass-trough)
	SF	B (Pass-trough)
	MS	B (Pass-trough)
	WTR	B (Pass-trough)
	EXER	B (Pass-trough)
	RR	N/A
	NR	N/A
Initial state	New request	New state
B (Pass-trough)	LP	B (Pass-trough)
	FS	N/A - cannot happen when there is LP request in the ring
	SF	B (Pass-trough) - otherwise N/A - cannot happen when there is LP request in the ring
	MS	B (Pass-trough) - otherwise N/A - cannot happen when there is LP, FS or SF request in the ring
	WTR	B (Pass-trough) - otherwise N/A - cannot happen when there is LP, FS, SF or MS request in the ring
	EXER	B (Pass-trough) - otherwise N/A - cannot happen when there is LP, FS, SF, MS or WTR request in the ring
	RR	B (Pass-trough) - otherwise N/A
	NR	B (Pass-trough)
Initial state	New request	New state
C (Switching - LP)	LP	C (Switching - LP)
	FS	N/A - cannot happen when there is LP request in the ring
	SF	N/A - cannot happen when there is LP request in the ring
	MS	N/A - cannot happen when there is LP request in the ring
	WTR	N/A - cannot happen when there is LP in the ring
	EXER	N/A - cannot happen when there

		is LP request in the ring
	RR	N/A
	NR	N/A
=====		
Initial state	New request	New state
-----	-----	-----
D (Idle - LW)	LP	B (Pass-trough)
	FS	B (Pass-trough)
	SF	B (Pass-trough)
	MS	B (Pass-trough)
	WTR	B (Pass-trough)
	EXER	B (Pass-trough)
	RR	N/A
	NR	N/A
=====		
Initial state	New request	New state
-----	-----	-----
E (Switching - FS)	LP	B (Pass-trough)
	FS	E (Switching - FS)
	SF	E (Switching - FS)
	MS	N/A - cannot happen when there is FS request in the ring
	WTR	N/A - cannot happen when there is FS request in the ring
	EXER	N/A - cannot happen when there is FS request in the ring
	RR	N/A
	NR	N/A
=====		
Initial state	New request	New state
-----	-----	-----
F (Switching - SF)	LP	B (Pass-trough)
	FS	F (Switching - SF)
	SF	F (Switching - SF)
	MS	N/A - cannot happen when there is SF request in the ring
	WTR	N/A - cannot happen when there is SF request in the ring
	EXER	N/A - cannot happen when there is SF request in the ring
	RR	N/A
	NR	N/A
=====		
Initial state	New request	New state
-----	-----	-----
G (Switching - MS)	LP	B (Pass-trough)
	FS	B (Pass-trough)
	SF	B (Pass-trough)

	MS	G (Switching - MS) - release the switches but signal MS
	WTR	N/A - cannot happen when there is MS request in the ring
	EXER	N/A - cannot happen when there is MS request in the ring
	RR	N/A
	NR	N/A
=====		
Initial state	New request	New state
-----	-----	-----
H (Switching - WTR)	LP	B (Pass-trough)
	FS	B (Pass-trough)
	SF	B (Pass-trough)
	MS	B (Pass-trough)
	WTR	N/A
	EXER	N/A - cannot happen when there is WTR request in the ring
	RR	N/A
	NR	N/A
=====		
Initial state	New request	New state
I (Switching - EXER)	LP	B (Pass-trough)
	FS	B (Pass-trough)
	SF	B (Pass-trough)
	MS	B (Pass-trough)
	WTR	N/A
	EXER	I (Switching - EXER)
	RR	N/A
	NR	N/A
=====		

4. IANA Considerations

Channel Types for the Generic Associated Channel are allocated from the IANA PW Associated Channel Type registry defined in RFC4446 [RFC4446] and updated by RFC5586 [RFC5586].

IANA is requested to allocate further Channel Type as follows:

- o 0xXX Ring Protection Switching (RPS) Note to RFC Editor:

this section may be removed on publication as an RFC.

5. Security Considerations

This document does not by itself raise any particular security considerations.

6. 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.
- [RFC4446] Martini, L., "IANA Allocations for Pseudowire Edge to Edge Emulation (PWE3)", BCP 116, RFC 4446, April 2006.
- [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.
- [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

Email: chengweiqiang@chinamobile.com

Lei Wang
China Mobile

Email: Wangleiyj@chinamobile.com

Han Li
China Mobile

Email: Lihan@chinamobile.com

Huub van Helvoort
Huawei Technologies Co., Ltd.

Email: huub.van.helvoort@huawei.com

Kai Liu
Huawei Technologies Co., Ltd.

Email: alex.liukai@huawei.com

Jia He
Huawei Technologies Co., Ltd.

Email: hejia@huawei.com

Fang Li
China Academy of Telecommunication Research, MIIT., China

Email: lifang@ritr.cn

Jian Yang
ZTE Corporation P.R.China

Email: yang.jian90@zte.com.cn

Junfang Wang
Fiberhome Telecommunication Technologies Co., LTD.

Email: wjf@fiberhome.com.cn

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: January 5, 2015

Z. Cui
R. Winter
NEC
H. Shah
Ciena
S. Aldrin
Huawei Technologies
M. Daikoku
KDDI
July 4, 2014

Use Cases and Requirements for MPLS-TP multi-failure protection
draft-cui-mpls-tp-mfp-use-case-and-requirements-02

Abstract

The basic survivability technique has been defined in Multiprotocol Label Switching Transport Profile (MPLS-TP) network [RFC6378]. That protocol however is limited to 1+1 and 1:1 protection, not designed to handle multi-failure protection.

This document introduces some use cases and requirements for multi-failure protection functionality.

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 5, 2015.

Copyright Notice

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

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

Table of Contents

1. Introduction	2
1.1. Document scope	3
1.2. Requirements notation	3
2. m:n protection architecture	3
3. Use cases	4
3.1. Increase service availability	4
3.2. Reduce the backup costs	5
4. Requirements	5
5. Security Considerations	5
6. IANA Considerations	6
7. Normative References	6
Authors' Addresses	6

1. Introduction

Today's packet optical transport networks are able to concentrate large volumes of traffic onto a relatively small number of nodes and links. As a result, the failure of a single network element can potentially interrupt a large amount of traffic. For this reason, ensuring survivability through fault-tolerant network design is an important network design objective.

The basic survivability technique has been defined in MPLS-TP network [RFC6378]. That protocol however is limited to 1+1 and 1:1 protection, not designed to handle multi-failure protection.

The case of multi-failure condition is very rare, but not unheard of. For example, when a working path was closed by network operator for construction work, the network service will become a hazardous condition. During this time, if another failure (e.g. a human-error or network entities failure) is occurred on the protection path, than the operator can't meet service level agreements (SLA).

A network must be able to handle multiple failures even that are a rare case, because especially some high-priority services such as emergency telephone calls request to network service provider

guarantee their service connections in a timely manner in any situation.

On the other hand, many network operators have a very limited budget for improving network survivability. This requires a design approach, which takes budget limitations into consideration.

To increase the service availability and to reduce the backup network costs, we propose extend the 1+1 and 1:1 protection protocol to support the m:n architecture type.

1.1. Document scope

This document describes the use cases and requirements for multi-failure protection in MPLS-TP networks without the use of control plane protocols. Existing solutions based on control plane such as GMPLS may be able to restore user traffic when multiple failures occur. Some networks however do not use full control plane operation for reasons such as service provider preferences, certain limitations or the requirement for fast service restoration (faster than achievable with control plane mechanisms). These networks are the focus of this document which defines a set of requirements for multi-failure protection not based on control plane support.

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. m:n protection architecture

The following Figure 1 shows a protection domain with n working paths and m protection paths. when a working path is determined to be impaired, its normal traffic must be assigned to a protection path if a protection path is available. To reduce the backup network costs, m protection paths are sharing backup resource for n working paths, where $m \leq n$ typically. The bandwidth of each protection paths should be allocated enough to protect any of the n working paths.

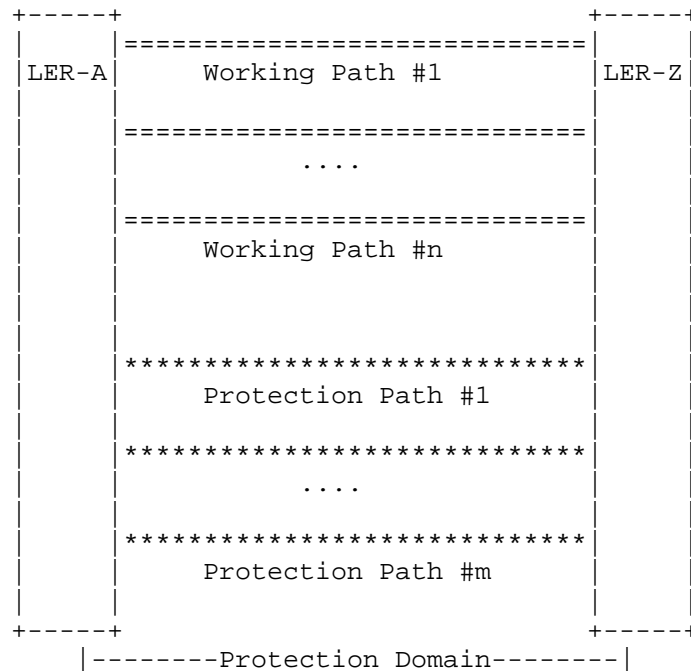


Figure 1: m:n ptorection domain

3. Use cases

3.1. Increase service availability

With technological advancement of mobile services or data center services, dependencies and business impact of network services have been increased phenomenally. End-users expectations of service availability also are increasing, which is driving service providers enhance their network's availability.

Network availability must be maintained especially for high-priority services such as emergency telephone calls, even during natural disasters and other catastrophic events such as earthquake or tsunami. Existing 1+1 or 1:n protection however is limited to cover single failure and no sufficient to maintain disaster recovery.

The m:n protection can increase service availability because it take multiple protection paths to ensuring high-priority services continue to operate on the 2nd, 3rd or Nth alternate backup, at least one of m protection paths is a available.

3.2. Reduce the backup costs

Network costs driven by high traffic growth rates are rising steadily, but revenues are not increased in direct proportion to traffic growth rates. This requires a design approach, which takes budget limitations into consideration.

Existing protection schemes such as 1+1 protection meet the sub 50 ms performance requirement but only protect against a single failure and are too costly.

The m:n protection is a useful solution, that can reduce the backup costs because m dedicated protection paths are sharing backup paths for n working paths, where $m \leq n$ typically.

The shared Mesh Protection (SMP) also can reduce the backup costs as described in [I-D.ietf-mpls-smp-requirements]. SMP however is based on the 1:1 protection and does not take care that the multiple failures are occurred on both working and protection paths. However, combine use of SMP and a set of m:1 protections to make a m:n protection likely, may be better able to recover the multiple failures.

4. Requirements

Some recovery requirements are defined [RFC5654]. That however is limited to cover single failure and is not able to take care that the multiple failures. This Section 4 extends the requirements to support the multiple failures scenarios.

MPLS-TP MUST support m:n protection with the following requirements:

- R1 The m:n protection MUST protect against multiple failures that are simultaneously-detected on both of working path and protection path or more than one multiple working paths.
- R2 Some priority schemes MUST be provided, because the backup resources are shared by multiple working paths dynamically.
- R3 TBD

5. Security Considerations

TBD

6. IANA Considerations

TBD

7. Normative References

- [I-D.ietf-mpls-smp-requirements]
Weingarten, Y., Aldrin, S., Pan, P., Ryoo, J., and G. Mirsky, "Requirements for MPLS-TP Shared Mesh Protection", draft-ietf-mpls-smp-requirements-06 (work in progress), June 2014.
- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC3945] Mannie, E., "Generalized Multi-Protocol Label Switching (GMPLS) Architecture", RFC 3945, October 2004.
- [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.

Authors' Addresses

Zhenlong Cui
NEC

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

Rolf Winter
NEC

Email: Rolf.Winter@neclab.eu

Himanshu Shah
Ciena

Email: hshah@ciena.com

Sam Aldrin
Huawei Technologies

Email: aldrin.ietf@gmail.com

Masahiro Daikoku
KDDI

Email: ms-daikoku@kddi.com

MPLS Working Group
INTERNET-DRAFT
Intended Status: Proposed Standard
Expires: October 4, 2014

Santosh Esale
Raveendra Torvi
Chris Bowers
Juniper Networks
April 2, 2014

Applications aware LDP Targeted Session
draft-esale-mpls-appl-aware-ldp-targeted-session-00

Abstract

Recent Targeted LDP applications such as Remote LFA and BGP auto discovery FEC 129 pseudowire may automatically establish a targeted LDP session to any LSR in the core network. The sender LSR has information about the targeted applications to administratively control initiation of the session. However the receiver LSR has no such information to control the acceptance of this session. This document defines a mechanism to advertise Targeted Application Capability during session initialization. As the receiver LSR becomes aware of targeted LDP applications, it may establish a limited number of sessions for certain applications. In addition, each targeted application is mapped to LDP FEC Elements to advertise only necessary LDP FEC label bindings over the session.

Status of this Memo

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

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

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

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

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

Copyright and License Notice

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

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

Table of Contents

1	Introduction	3
1.1	Terminology	3
2.	Targeted Application Capability	4
3.	Targeted Application Capability Procedures	5
4.	Interaction of Targeted Application Capabilities and State Advertisement Control Capabilities	6
5.	Targeted Application capability in LDP messages	8
5.1	TAC in LDP Initialization message	8
5.2	TAC in LDP Capability message	8
7.	Use cases	8
7.1	Remote LFA Automatic Targeted session	8
7.2	FEC 129 Auto Discovery Targeted session	9
7.3	LDP over RSVP and Remote LFA targeted session	9
8	Security Considerations	10
9	IANA Considerations	10
10.	Acknowledgments	11
11	References	11
11.1	Normative References	11
11.2	Informative References	11
	Authors' Addresses	12

1 Introduction

LDP can use the extended discovery mechanism to establish a targeted adjacency and subsequent session as described in [RFC5036]. An LSR initiates extended discovery by sending the targeted Hello to a specific address. The remote LSR decides either to accept or ignore the Hello based on local configuration only. For an application such as FEC 128 pseudowire and LDP over RSVP tunneling, the remote LSR is configured with the source LSR address, so the remote LSR can use that information to accept or ignore any given LDP Hello.

Applications such as remote LFA and FEC 129 pseudowire automatically initiate asymmetric extended discovery to any LSR in the network based on local state. In these applications, the remote LSR is not explicitly configured with the source LSR address, so the remote LSR either responds to all LDP requests or ignores all LDP requests.

In addition, since the session is initiated and established after adjacency formation, the receiver LSR has no targeted application information to choose the targeted applications it would like to support. While the sender LSR may employ a limit per application on locally initiated automatic targeted sessions, the receiver LSR currently has no mechanism to apply a similar limit on the incoming targeted sessions. Also, the receiver LSR does not know whether the source LSR is establishing the session for a configured or an automatic application.

This document proposes and describes a solution to advertise targeted application capability, consisting of a targeted application list, during initialization of a targeted session. It also defines a mechanism to enable a new application and disable an old application after session establishment. This capability advertisement provides the remote LSR with the necessary information to control the targeted sessions per application. For instance, an LSR may wish to accept all FEC 129 targeted sessions but may only accept limited number of Remote LFA targeted sessions.

Also, targeted applications may be mapped to LDP FEC type to advertise specific application FECs only, avoiding the advertisement of unnecessary FECs over the session.

1.1 Terminology

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

2. Targeted Application Capability

An LSR MAY advertise that it is capable of negotiating a targeted application list over a session by using the Capability Advertisement as defined in [RFC5561].

A new optional capability TLV is defined, 'Targeted Application Capability (TAC)'. Its encoding is as follows:

```

      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| Targeted App. Cap.(IANA)|                               Length|
+-----+-----+-----+-----+-----+-----+-----+-----+
|S|   Reserved   |                                           |
+-----+-----+-----+-----+-----+-----+-----+-----+
|                                                           |
~                               Targeted App. Cap. data      ~
|                                                           |
+-----+-----+-----+-----+-----+-----+-----+-----+

```

U: set to 1. Ignore, if not known.

F: Set to 0. Do not forward.

S: MUST be set to 1 and ignored on receipt.

Targeted Application Capability data:

A Targeted Applications Capability data consists of none, one or more Targeted Application Elements. Its encoding is as follows:

Targeted Application Element(TAE)

```

      0               1
      0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5
+-----+-----+-----+-----+-----+-----+-----+
| Targ. Appl. Id|E|   Reserved   |
+-----+-----+-----+-----+-----+-----+-----+

```

Targeted Application Identifier (TA-Id):

```

0x0001: LDPv4 Tunneling
0x0002: LDPv6 Tunneling
0x0003: mLDP Tunneling
0x0004: LDPv4 Remote LFA
0x0005: LDPv6 Remote LFA
0x0006: FEC 128 Pseudowire
0x0007: FEC 129 Pseudowire

```

E-bit: It indicates whether the sender is advertising or withdrawing the Targeted Application. The E-bit value is used as follows:

- 1 - The TAE is advertising the targeted application.
- 0 - The TAE is withdrawing the targeted application.

The length of TAC depends on the number of TAE elements. For instance, if two TAE elements are added, the length is set to 5.

If both the peers advertise TAC, an LSR decides to establish or close a targeted session based on the negotiated targeted application element list.

3. Targeted Application Capability Procedures

At targeted session establishment time, an LSR MAY include a new capability TLV, Targeted Application Capability (TAC) TLV, as an optional TLV in the LDP Initialization message. The TAC TLV's Capability data MUST consists of none, one or more Targeted Application Element(TAE) each pertaining to a unique Targeted Application Identifier(TA-Id). If the receiver LSR receives the same TA-Id in more than one TAE element, it MUST discard the TAC TLV and behave as if TAC TLV is not received. If the receiver LSR receives an unknown TA-Id in a TAE element, it SHOULD silently ignore such a TAE element and continue processing the rest of the TLV.

If the receiver LSR does not receive the TAC in the Initialization message or it does not understand the TAC TLV, the TAC negotiation MUST be considered unsuccessful and the session establishment MUST proceed as per [RFC5036]. On the receipt of a valid TAC TLV, an LSR MUST generate its own TAC TLV with TAE elements consisting of unique TA-Ids that it MAY support over the targeted session. If there is at least one TAE element common between the TAC TLV it has received and its own, the session MUST proceed to establishment as per [RFC5036]. If not, A LSR MUST send a 'Session Rejected/Targeted Application Capability Mis-Match' Notification message to the peer and close the session. The sender LSR playing the passive role in LDP session establishment MAY destroy the corresponding targeted adjacency.

When the receiver LSR playing the active role in LDP session establishment receives a 'Session Rejected/Targeted Application Capability Mis-Match' Notification message, it MUST set its session setup retry interval to a maximum value, as 0xffff. The session MAY stay in non-operational state. When it detects a change in the sender LSR configuration or local configuration pertaining to TAC TLV, it

MUST clear the session setup back off delay associated with the session to re-attempt the session establishment.

When the sender LSR playing the active role in LDP session establishment receives a 'Session Rejected/Targeted Application Capability Mis-Match' Notification message, either it MUST set its session setup retry interval to a maximum value, as 0xffff or it MUST destroy the corresponding targeted adjacency with the session. This leads to destruction of the session.

If it sets the session setup retry interval to maximum, the session MAY stay in a non-operational state. When this LSR detects a change in the receiver LSR configuration or its own configuration pertaining to TAC TLV, it MUST clear the session setup back off delay associated with the session to re-attempt the session establishment.

If it decides to destroy the associated targeted adjacency, the session is destroyed on the sender as well as the receiver LSR. The sender LSR MAY take appropriate actions if it is unable to bring up the targeted session. For instance, if an automatic session intended to support the Remote LFA application is rejected by the receiver LSR, the sender LSR MAY inform the IGP to calculate another PQ node for the route or set of routes. More specific actions are a local matter and outside the scope of this document.

After an LDP session has been established with TAC capability, the sender and receiver LSR MUST distribute FEC label bindings for the negotiated applications only. For instance, if the LDP session is established for FEC 129 pseudowire, only FEC 129 label bindings MUST be distributed over the session. Similarly, a LSR MUST request FEC label bindings for the negotiated applications only.

If the Targeted Application Capability and Dynamic Capability, as defined in RFC5561, are negotiated during session Initialization, TAC MAY be re-negotiated after session establishment by sending the updated TAC TLV in LDP Capability message. The Updated TLV MUST consist of one or more TAE elements with E bit set or E bit off to advertise or withdraw the new and old application respectively. This MAY lead to advertisements or withdrawals of certain FEC types over the session or destruction of the adjacency and subsequently the session.

4. Interaction of Targeted Application Capabilities and State Advertisement Control Capabilities

As described in this document, the set of Targeted Application Elements negotiated between two LDP peers advertising TAC represents the willingness of both peers to advertise state information for a

set of applications. The set of applications negotiated by the TAC mechanism is symmetric between the two LDP peers. In the absence of further mechanisms, two LDP peers will both advertise state information for the same set of applications.

As described in [I-D.draft-ietf-mpls-ldp-ip-pw-capability], State Advertisement Control(SAC) TLV can be used by an LDP speaker to communicate its interest or disinterest in receiving state information from a given peer for a particular application. Two LDP peers can use the SAC mechanism to create asymmetric advertisement of state information between the two peers for any particular application.

For a given LDP session, the TAC mechanism can be used without the SAC mechanism, and the SAC mechanism can be used without the TAC mechanism. It is useful to discuss the behavior when TAC and SAC mechanisms are used on the same LDP session. The TAC mechanism takes precedence over the SAC mechanism with respect to enabling applications for which state information will be advertised. For an LDP session using the TAC mechanism, the LDP peers MUST NOT advertise state information for an application that has not been negotiated in the most recent Targeted Application Elements list (referred to as an un-negotiated application). This is true even if one of the peers announces its interest in receiving state information that corresponds to the un-negotiated application by sending a SAC TLV. In other words, when TAC is being used, SAC cannot enable state information advertisement for applications that have not been enabled by TAC.

On the other hand, the SAC mechanism takes precedence over the TAC mechanism with respect to disabling state information advertisements. If an LDP speaker has announced its disinterest in receiving state information for a given application to a given peer using the SAC mechanism, its peer MUST NOT send state information for that application, even if the two peers have negotiated that the corresponding application via the TAC mechanism.

For the purposes of determining the correspondence between targeted applications defined in this document and application state as defined in [I-D.draft-ietf-mpls-ldp-ip-pw-capability] an LSR MAY use the following mappings:

- LDPv4 Tunneling - IPv4 Label switching
- LDPv6 Tunneling - IPv6 Label switching
- LDPv4 Remote LFA - IPv4 Label switching
- LDPv6 Remote LFA - IPv6 Label switching
- FEC 128 Pseudowire - P2P PW FEC128 signaling
- FEC 129 Pseudowire - P2P PW FEC129 signaling

An LSR MAY map Targeted Application to LDP capability as follows:

mLDP Tunneling - P2MP Capability, MP2MP Capability

5. Targeted Application capability in LDP messages

5.1 TAC in LDP Initialization message

1. The S-bit of the Targeted Application Capability parameter MUST be set to 1 to advertise Targeted Application Capability and SHOULD be ignored on the receipt. The E-bit of the Targeted Application Element MUST be set to 1 to enable Targeted application.
2. An LSR MAY add State Control Capability by mapping Targeted Application element to State Advertisement Control (SAC) Elements as defined in Section 4.
3. The LSR MAY add a different Hold time value for an automatic targeted session.

5.2 TAC in LDP Capability message

1. The S-bit of Targeted Application Capability is set to 1 and ignored on receipt.
2. If there is no common Targeted Application element between its new TAC and peers TAC, the LSR MUST send a 'Session Rejected/Targeted Application Capability Mis-Match' Notification message and close the session.
3. If there is a common Targeted Application Element, a LSR MAY also update State Advertisement Control Capability as per Section 4 and send these capabilities in a Capability message to the peer.
4. A receiving LSR processes the Capability message and its Targeted Applications Capability TLV. The S-bit is ignored on receipt.
5. Process a List of Targeted Application Elements from capability data with E-bit set to 1 to construct peers Targeted Application Capability.

7. Use cases

7.1 Remote LFA Automatic Targeted session

An LSR determines that it needs to form a automatic targeted session

to remote PQ node based on IGP calculation as described in [I-D.draft-ietf-rtgwg-remote-lfa] or some other mechanism, which is outside the scope of this document. The LSR forms the targeted adjacency and during session setup, constructs an Initialization message with Targeted Applications Capability (TAC) with Targeted Application Element (TAE) as Remote LFA. The receiver LSR processes the LDP Initialization message and verifies whether it is configured to accept a Remote LFA targeted session. If it is, it MAY further verify that establishing such a session does not exceed the configured limit for Remote LFA sessions. If all these conditions are met, the receiver LSR may respond back with an Initialization message with TAC corresponding to Remote LFA, and subsequently the session may be established.

After the session has been established with TAC capability, the sender and receiver LSR distribute IPv4 or IPv6 FEC label bindings over the session. Further, the receiver LSR may determine that it does not need these FEC label bindings. So it may disable the receipt of these FEC label bindings by mapping targeted application element to state control capability as described in section 4.

7.2 FEC 129 Auto Discovery Targeted session

BGP auto discovery or other mechanisms outside the scope of this document MAY determine whether an LSR needs to initiate an auto-discovery targeted session with a border LSR. Multiple LSRs MAY try to form an auto-discovery LDP targeted session with a border LSR. So a service provider may want to limit the number of auto-discovery targeted sessions a border LSR MAY accept. As described in Section 3, LDP MAY convey Targeted Applications with TAC TLV to border LSR. A border LSR may establish or reject the session based on local administrative policy. Also, as the receiver LSR would be aware of targeted application, it can also employ an administrative policy for security. For instance, it can employ a policy 'accept all auto-discovered session from source-list'.

Moreover, the sender and receiver LSR MUST exchange FEC 129 application states only over the targeted session, i.e. FEC 129 label bindings only.

7.3 LDP over RSVP and Remote LFA targeted session

A LSR may want to establish a targeted session to a remote LSR for LDP over RSVP tunneling and Remote LFA. The sender LSR may add both these applications as a unique Targeted Application Element in the Targeted Application Capability data of a TAC TLV. The receiver LSR MAY have reached a configured limit for accepting automatic targeted sessions for Remote LFA, but it may also be configured to accept LDP

over RSVP tunneling. In this case, the targeted session is formed for both LDP over RSVP and Remote LFA applications as both needs same FECs - IPv4 and/or IPv6.

Also, the sender and the receiver LSR MUST distributes IPv4 and or IPv6 FEC label bindings only over the session.

8 Security Considerations

The Capability procedure described in this document will apply and does not introduce any change to LDP Security Considerations section described in [RFC5036].

9 IANA Considerations

This document requires the assignment of a new code point for a Capability Parameter TLVs from the IANA managed LDP registry "TLV Type Name Space", corresponding to the advertisement of the Targeted Applications capability. IANA is requested to assign the lowest available value after 0x050B.

Value	Description	Reference
-----	-----	-----
TBD1	Targeted Applications capability	[This draft]

This document requires the assignment of a new code point for a status code from the IANA managed registry "STATUS CODE NAME SPACE", corresponding to the notification of session Rejected/Targeted Application Capability Mis-Match. IANA is requested to assign the lowest available value after 0x0000004B.

Value	Description	Reference
-----	-----	-----
TBD2	Session Rejected/Targeted Application Capability Mis-Match	[This draft]

This document also creates a new name space 'the LDP Targeted Application Element type' that is to be managed by IANA. The range is 0-65535, with the following values requested in this document.

```

0x0000: Reserved
0x0001: LDPv4 Tunneling
0x0002: LDPv6 Tunneling
0x0003: mLDP Tunneling
0x0004: LDPv4 Remote LFA
0x0005: LDPv6 Remote LFA

```

0x0006: FEC 128 Pseudowire
0x0007: FEC 129 Pseudowire

The allocation policy for this space is 'Standards Action with Early Allocation'.

10. Acknowledgments

The authors wish to thank Nischal Sheth, Hassan Hosseini and Kishore Tiruveedhula for doing the detailed review. Thanks to Manish Gupta and Martin Ehlers for their input to this work and for many helpful suggestions.

11. References

11.1 Normative References

- [RFC5036] Andersson, L., Ed., Minei, I., Ed., and B. Thomas, Ed., "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.
- [I-D.draft-ietf-mppls-ldp-ip-pw-capability] Kamran Raza, Sami Boutros, "Disabling IPoMPLS and P2P PW LDP Application's State Advertisement", draft-ietf-mppls-ldp-ip-pw-capability-06 (work in progress), December 19, 2013.
- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997..

11.2 Informative References

- [I-D.draft-ietf-rtgwg-remote-lfa] S. Bryant, C. Filsfils, S. Previdi, M. Shand, "Remote LFA FRR", draft-ietf-rtgwg-remote-lfa-04 (work in progress), November 22, 2013.
- [RFC6074] E. Rosen, B. Davie, V. Radoaca, and W. Luo, "Provisioning, Auto-Discovery, and Signaling in Layer 2 Virtual Private Networks (L2VPNs)"
- [RFC4762] M. Lasserre, and V. Kompella, "Virtual Private LAN Service (VPLS) Using Label Distribution Protocol (LDP) Signaling", RFC 4762, January 2007.
- [RFC4447] L. Martini, E. Rosen, El-Aawar, T. Smith, and G. Heron,

"Pseudowire Setup and Maintenance using the Label
Distribution Protocol", RFC 4447, April 2006.

[RFC5331] Aggarwal, R., Rekhter, Y., and E. Rosen, "MPLS Upstream
Label Assignment and Context-Specific Label Space", RFC
5331, August 2008.

Authors' Addresses

Santosh Esale
Juniper Networks
1194 N. Mathilda Ave.
Sunnyvale, CA 94089
US
EMail: sesale@juniper.net

Raveendra Torvi
Juniper Networks
10 Technology Park Drive.
Westford, MA 01886
US
EMail: rtorvi@juniper.net

Chris Bowers
Juniper Networks
1194 N. Mathilda Ave.
Sunnyvale, CA 94089
US
EMail: cbowers@juniper.net

Internet Engineering Task Force
Internet-Draft
Intended status: Standards Track
Expires: January 4, 2015

H. Chen
Z. Li
Huawei Technologies
N. So
Tata Communications
A. Liu
Ericsson
F. Xu
Verizon
M. Toy
Comcast
L. Huang
China Mobile
L. Liu
UC Davis
July 3, 2014

Extensions to RSVP-TE for LSP Egress Local Protection
draft-ietf-mpls-rsvp-egress-protection-01.txt

Abstract

This document describes extensions to Resource Reservation Protocol - Traffic Engineering (RSVP-TE) for locally protecting egress nodes of a Traffic Engineered (TE) Label Switched Path (LSP) in a Multi-Protocol Label Switching (MPLS) and Generalized MPLS (GMPLS) network.

Status of this Memo

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

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

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

This Internet-Draft will expire on January 4, 2015.

Copyright Notice

Copyright (c) 2014 IETF Trust and the persons identified as the

document authors. All rights reserved.

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

Table of Contents

1. Introduction	3
1.1. An Example of Egress Local Protection	3
1.2. Egress Local Protection with FRR	4
2. Conventions Used in This Document	4
3. Terminology	4
4. Protocol Extensions	4
4.1. EGRESS_BACKUP Object	4
4.2. Flags in FAST_REROUTE	6
4.3. Path Message	6
5. Egress Protection Behaviors	6
5.1. Ingress Behavior	6
5.2. Intermediate Node and PLR Behavior	7
5.2.1. Signaling for One-to-One Protection	8
5.2.2. Signaling for Facility Protection	8
5.2.3. Signaling for S2L Sub LSP Protection	9
5.2.4. PLR Procedures during Local Repair	10
6. Considering Application Traffic	10
6.1. A Typical Application	10
6.2. PLR Procedure for Applications	11
6.3. Egress Procedures for Applications	11
7. Security Considerations	12
8. IANA Considerations	12
9. Contributors	12
10. Acknowledgement	13
11. References	13
11.1. Normative References	13
11.2. Informative References	14
Authors' Addresses	14

1. Introduction

RFC 4090 describes two methods for protecting the transit nodes of a P2P LSP: one-to-one and facility protection. RFC 4875 specifies how to use them to protect the transit nodes of a P2MP LSP. However, they do not mention any local protection for an egress of an LSP.

To protect the egresses of an LSP (P2P or P2MP), an existing approach sets up a backup LSP from a backup ingress (or the ingress of the LSP) to the backup egresses, where each egress is paired with a backup egress and protected by the backup egress.

This approach may use more resources and provide slow fault recovery. This document specifies extensions to RSVP-TE for local protection of an egress of an LSP, which overcomes these disadvantages.

1.1. An Example of Egress Local Protection

Figure 1 shows an example of using backup LSPs to locally protect egresses of a primary P2MP LSP from ingress R1 to two egresses: L1 and L2. The primary LSP is represented by star(*) lines and backup LSPs by hyphen(-) lines.

La and Lb are the designated backup egresses for egresses L1 and L2 respectively. To distinguish an egress (e.g., L1) from a backup egress (e.g., La), an egress is called a primary egress if needed.

The backup LSP for protecting L1 is from its upstream node R3 to backup egress La. The one for protecting L2 is from R5 to Lb.

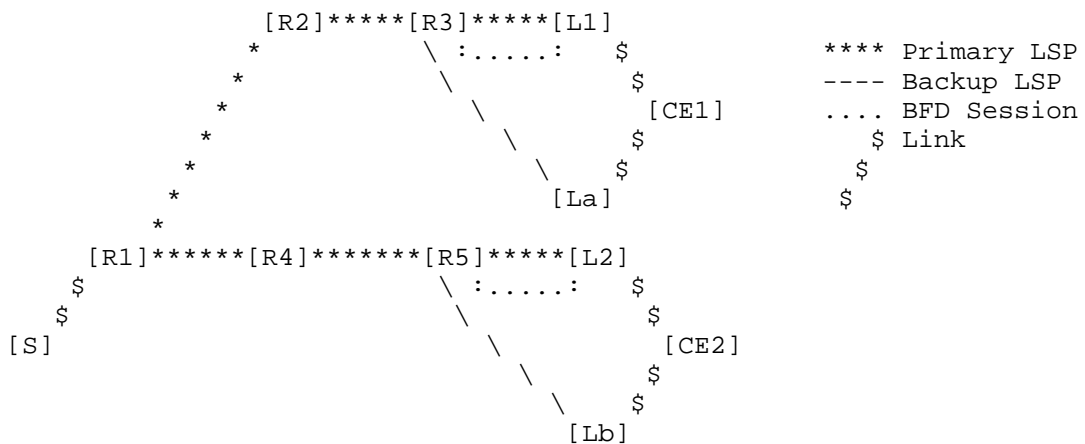


Figure 1: Backup LSP for Locally Protecting Egress

During normal operations, the traffic carried by the P2MP LSP is sent through R3 to L1, which delivers the traffic to its destination CE1. When R3 detects the failure of L1, R3 switches the traffic to the backup LSP to backup egress La, which delivers the traffic to CE1. The time for switching the traffic is within tens of milliseconds.

The failure of a primary egress (e.g., L1 in the figure) MAY be detected by its upstream node (e.g., R3 in the figure) through a BFD between the upstream node and the egress in MPLS networks. Exactly how the failure is detected is out of scope for this document.

1.2. Egress Local Protection with FRR

Using the egress local protection and the FRR, we can locally protect the egresses, the links and the intermediate nodes of an LSP. The traffic switchover time is within tens of milliseconds whenever an egress, any of the links and the intermediate nodes of the LSP fails.

The egress nodes of the LSP can be locally protected via the egress local protection. All the links and the intermediate nodes of the LSP can be locally protected through using the FRR.

2. Conventions Used in This Document

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

3. Terminology

This document uses terminologies defined in RFC 2205, RFC 3031, RFC 3209, RFC 3473, RFC 4090, RFC 4461, and RFC 4875.

4. Protocol Extensions

A new object EGRESS_BACKUP is defined for egress local protection. It contains a backup egress for a primary egress.

4.1. EGRESS_BACKUP Object

The class of the EGRESS_BACKUP object is TBD-1 to be assigned by IANA. The C-Type of the EGRESS_BACKUP IPv4/IPv6 object is TBD-2/TBD-3 to be assigned by IANA.

EGRESS_BACKUP Class Num = TBD-1, IPv4/IPv6 C-Type = TBD-2/TBD-3

```

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
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
~                               Backup Egress IPv4/IPv6 address      ~
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
~                               Primary Egress IPv4/IPv6 address      ~
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
~                               (Subobjects)                          ~
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+

```

- o Backup Egress IPv4/IPv6 address:
IPv4/IPv6 address of the backup egress node
- o Primary Egress IPv4/IPv6 address:
IPv4/IPv6 address of the primary egress node

The Subobjects are optional. One of them is P2P LSP ID IPv4/IPv6 subobject, whose body has the following format and Type is TBD-4/TBD-5. It may be used to identify a backup LSP.

```

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
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
~                               P2P LSP Tunnel Egress IPv4/IPv6 Address (4/16 bytes) ~
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
|                               Reserved                               | Tunnel ID |
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
~                               Extended Tunnel ID (4/16 bytes)      ~
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+

```

- o P2P LSP Tunnel Egress IPv4/IPv6 Address:
IPv4/IPv6 address of the egress of the tunnel
- o Tunnel ID:
A 16-bit identifier that is constant over the life of the tunnel
- o Extended Tunnel ID:
A 4/16-byte identifier being constant over the life of the tunnel

Another one is Label subobject, whose body has the format below and Type is TBD-6 to be assigned by IANA.

```

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
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
|                               Label                               |
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+

```

4.2. Flags in FAST_REROUTE

A bit of the flags in the FAST_REROUTE object may be used to indicate whether S2L Sub LSP is desired for protecting an egress of a P2MP LSP or One-to-One Backup is preferred for protecting an egress of a P2P LSP when the "Facility Backup Desired" flag is set. This bit is called "S2L Sub LSP Backup Desired" or "One-to-One Backup Preferred".

4.3. Path Message

A Path message is enhanced to carry the information about a backup egress for a primary egress of an LSP through including an egress backup descriptor list. The format of the enhanced Path message is illustrated below.

```
<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> [ <S2L sub-LSP descriptor list> ]
                    [ <egress backup descriptor list> ]
```

The egress backup descriptor list in the message is defined below. It is a sequence of EGRESS_BACKUP objects, each of which describes a pair of a primary egress and a backup egress.

```
<egress backup descriptor list> ::=
    <egress backup descriptor>
    [ <egress backup descriptor list> ]

<egress backup descriptor> ::= <EGRESS_BACKUP>
```

5. Egress Protection Behaviors

5.1. Ingress Behavior

To protect a primary egress of an LSP, the ingress MUST set the "label recording desired" flag and the "node protection desired" flag in the SESSION_ATTRIBUTE object.

If one-to-one backup or facility backup method is desired to protect a primary egress of an LSP, the ingress SHOULD include a FAST_REROUTE

object and set the "One-to-One Backup Desired" or "Facility Backup Desired" flag.

If S2L Sub LSP backup method is desired to protect a primary egress of a P2MP LSP, the ingress SHOULD include a FAST_REROUTE object and set the "S2L Sub LSP Backup Desired" flag.

Note that if "Facility Backup Desired" flag is set for protecting the intermediate nodes of a primary P2P LSP, but we want to use "One-to-One Backup" for protecting the egress of the LSP, then the ingress SHOULD set "One-to-One Backup Preferred" flag.

Optionally, a backup egress may be configured on the ingress of an LSP to protect a primary egress of the LSP.

The ingress sends a Path message for the LSP with the objects above and an optional egress backup descriptor list. For each primary egress of the LSP to be protected, the ingress adds an EGRESS_BACKUP object into the list if the backup egress is given. The object contains the primary egress and the backup egress for protecting the primary egress.

5.2. Intermediate Node and PLR Behavior

If an intermediate node of an LSP receives the Path message with an egress backup descriptor list and it is not an upstream node of any primary egress of the LSP, it forwards the list unchanged.

If the intermediate node is the upstream node of a primary egress to be protected, it determines the backup egress, obtains a path for the backup LSP and sets up the backup LSP along the path.

The PLR (upstream node of the primary egress) tries to get the backup egress from EGRESS_BACKUP in the egress backup descriptor list if the Path message contains the list. If the PLR can not get it, the PLR tries to find the backup egress, which is not the primary egress but has the same IP address as the destination IP address of the LSP.

Note that the primary egress and the backup egress SHOULD have a same local address configured, and the cost to the local address on the backup egress SHOULD be much bigger than the cost to the local address on the primary egress. Thus another name such as virtual node based egress protection may be used for egress local protection.

After obtaining the backup egress, the PLR tries to compute a backup path from itself to the backup egress. It excludes the primary egress to be protected when computing the path. Thus the PLR will not select any path via the primary egress.

The PLR then sets up the backup LSP along the path obtained. It provides one-to-one backup protection for the primary egress if the "One-to-One Backup Desired" or "One-to-One Backup Preferred" flag is set in the message; otherwise, it provides facility backup protection if the "Facility Backup Desired flag" is set.

The PLR sets the protection flags in the RRO Sub-object for the primary egress in the Resv message according to the status of the primary egress and the backup LSP protecting the primary egress. For example, it will set the "local protection available" and the "node protection" flag indicating that the primary egress is protected when the backup LSP is up and ready for protecting the primary egress.

5.2.1. Signaling for One-to-One Protection

The behavior of the upstream node of a primary egress of an LSP as a PLR is the same as that of a PLR for one-to-one backup method described in RFC 4090 except for that the upstream node creates a backup LSP from itself to a backup egress.

If the LSP is a P2MP LSP and a primary egress of the LSP is also a transit node (i.e., bud node), the upstream node of the primary egress as a PLR also creates a backup LSP from itself to each of the next hops of the primary egress.

When the PLR detects the failure of the primary egress, it MUST switch the packets from the primary LSP to the backup LSP to the backup egress. For the failure of the bud node of a P2MP LSP, the PLR MUST also switch the packets to the backup LSPs to the bud node's next hops, where the packets are merged into the primary LSP.

5.2.2. Signaling for Facility Protection

Except for backup LSP and downstream label, the behavior of the upstream node of the primary egress of a primary LSP as a PLR follows the PLR behavior for facility backup method described in RFC 4090.

For a number of primary P2P LSPs going through the same PLR to the same primary egress, the primary egress of these LSPs may be protected by one backup LSP from the PLR to the backup egress designated for protecting the primary egress.

The PLR selects or creates a backup LSP from itself to the backup egress. If there is a backup LSP that satisfies the constraints given in the Path message, then this one is selected; otherwise, a new backup LSP to the backup egress will be created.

After getting the backup LSP, the PLR associates the backup LSP with

a primary LSP for protecting its primary egress. The PLR records that the backup LSP is used to protect the primary LSP against its primary egress failure and includes an EGRESS_BACKUP object in the Path message to the primary egress. The object contains the backup egress and the backup LSP ID. It indicates that the primary egress SHOULD send the backup egress the primary LSP label as UA label.

After receiving the Path message with the EGRESS_BACKUP, the primary egress includes the information about the primary LSP label in the Resv message with an EGRESS_BACKUP object as UA label. When the PLR receives the Resv message with the information about the UA label, it includes the information in the Path message for the backup LSP to the backup egress. Thus the primary LSP label as UA label is sent to the backup egress from the primary egress.

When the PLR detects the failure of the primary egress, it redirects the packets from the primary LSP into the backup LSP to backup egress using the primary LSP label from the primary egress as an inner label. The backup egress delivers the packets to the same destinations as the primary egress using the backup LSP label as context label and the inner label as UA label.

5.2.3. Signaling for S2L Sub LSP Protection

The S2L Sub LSP Protection is used to protect a primary egress of a P2MP LSP. Its major advantage is that the application traffic carried by the LSP is easily protected against the egress failure.

The PLR determines to protect a primary egress of a P2MP LSP via S2L sub LSP protection when it receives a Path message with flag "S2L Sub LSP Backup Desired" set.

The PLR sets up the backup S2L sub LSP to the backup egress, creates and maintains its state in the same way as of setting up a source to leaf (S2L) sub LSP defined in RFC 4875 from the signaling's point of view. It computes a path for the backup LSP from itself to the backup egress, constructs and sends a Path message along the path, receives and processes a Resv message responding to the Path message.

After receiving the Resv message for the backup LSP, the PLR creates a forwarding entry with an inactive state or flag called inactive forwarding entry. This inactive forwarding entry is not used to forward any data traffic during normal operations.

When the PLR detects the failure of the primary egress, it changes the forwarding entry for the backup LSP to active. Thus, the PLR forwards the traffic to the backup egress through the backup LSP, which sends the traffic to its destination.

5.2.4. PLR Procedures during Local Repair

When the upstream node of a primary egress of an LSP as a PLR detects the failure of the primary egress, it follows the procedures defined in section 6.5 of RFC 4090. It SHOULD notify the ingress about the failure of the primary egress in the same way as a PLR notifies the ingress about the failure of an intermediate node.

Moreover, the PLR lets the upstream part of the primary LSP stay after the primary egress fails. It continues to send resv message to its upstream node along the primary LSP. The downstream part of the primary LSP from the PLR to the primary egress SHOULD be removed.

In the local revertive mode, the PLR re-signals each of the primary LSPs that were routed over the restored resource once it detects that the resource is restored. Every primary LSP successfully re-signaled along the restored resource is switched back.

6. Considering Application Traffic

This section focuses on the application traffic carried by P2P LSPs. When a primary egress of a P2MP LSP fails, the application traffic carried by the P2MP LSP is delivered to the same destination by the backup egress since the inner label if any for the traffic is a upstream assigned label for every egress of the P2MP LSP.

6.1. A Typical Application

L3VPN is a typical application. An existing solution (refer to Figure 2) for protecting L3VPN traffic against egress failure includes: 1) A multi-hop BFD session between ingress R1 and egress L1 of primary LSP; 2) A backup LSP from ingress R1 to backup egress La; 3) La sends R1 VPN backup label and related information via BGP; 4) R1 has a VRF with two sets of routes: one uses primary LSP and L1 as next hop; the other uses backup LSP and La as next hop.

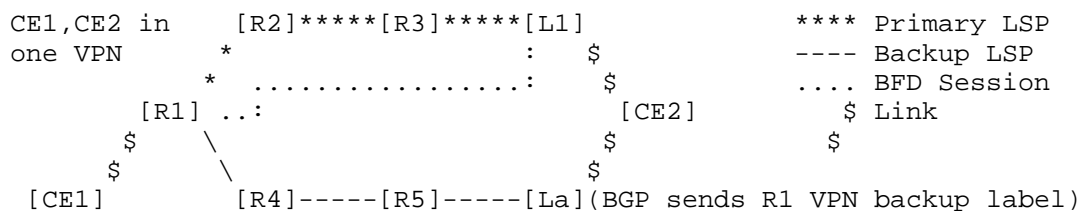


Figure 2: Protect Egress for L3VPN Traffic

In normal operations, R1 sends the traffic from CE1 through primary LSP with VPN label received from L1 as inner label to L1, which delivers the traffic to CE2 using VPN label.

When R1 detects the failure of L1, R1 sends the traffic from CE1 via backup LSP with VPN backup label received from La as inner label to La, which delivers the traffic to CE2 using VPN backup label.

A new solution (refer to Figure 3) with egress local protection for protecting L3VPN traffic includes: 1) A BFD session between R3 and egress L1 of primary LSP; 2) A backup LSP from R3 to backup egress La; 3) L1 sends La VPN label as UA label and related information; 4) L1 and La is virtualized as one. This can be achieved by configuring a same local address on L1 and La, using the address as a destination of the LSP and BGP next hop for VPN traffic.

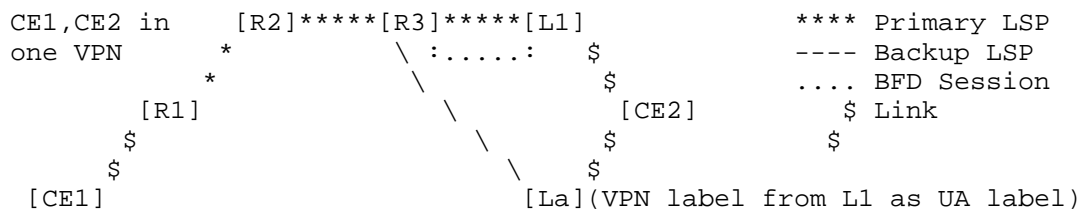


Figure 3: Locally Protect Egress for L3VPN Traffic

When R3 detects L1's failure, R3 sends the traffic from primary LSP via backup LSP to La, which delivers the traffic to CE2 using VPN label as UA label under the backup LSP label as a context label.

6.2. PLR Procedure for Applications

When the PLR gets a backup LSP from itself to a backup egress for protecting a primary egress of a primary LSP, it includes an EGRESS_BACKUP object in the Path message for the primary LSP. The object contains the ID information of the backup LSP and indicates that the primary egress SHOULD send the backup egress the application traffic label (e.g., VPN label) as UA label when needed.

6.3. Egress Procedures for Applications

When a primary egress of an LSP sends the ingress of the LSP a label for an application such as a VPN, it SHOULD send the backup egress for protecting the primary egress the label as a UA label via BGP or another protocol. Exactly how the label is sent is out of scope for this document.

When the backup egress receives a UA label from the primary egress, it adds a forwarding entry with the label into the LFIB for the primary egress. When the backup egress receives a packet from the backup LSP, it uses the top label as a context label to find the LFIB for the primary egress and the inner label to deliver the packet to the same destination as the primary egress according to the LFIB.

7. Security Considerations

In principle this document does not introduce new security issues. The security considerations pertaining to RFC 4090, RFC 4875 and other RSVP protocols remain relevant.

8. IANA Considerations

IANA considerations for new objects will be specified after the objects used are decided upon.

9. Contributors

Boris Zhang
Telus Communications
200 Consilium Pl Floor 15
Toronto, ON M1H 3J3
Canada
Email: Boris.Zhang@telus.com

Nan Meng
Huawei Technologies
Huawei Bld., No.156 Beijing Rd.
Beijing 100095
China
Email: mengnan@huawei.com

Vic Liu
China Mobile
No.32 Xuanwumen West Street, Xicheng District
Beijing, 100053
China
Email: liuzhiheng@chinamobile.com

10. Acknowledgement

The authors would like to thank Richard Li, Nobo Akiya, Tarek Saad, Lizhong Jin, Ravi Torvi, Eric Gray, Olufemi Komolafe, Michael Yue, Rob Rennison, Neil Harrison, Kannan Sampath, Yimin Shen, Ronhazli Adam and Quintin Zhao for their valuable comments and suggestions on this draft.

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.
- [RFC3692] Narten, T., "Assigning Experimental and Testing Numbers Considered Useful", BCP 82, RFC 3692, January 2004.
- [RFC2205] Braden, B., Zhang, L., Berson, S., Herzog, S., and S. Jamin, "Resource ReSerVation Protocol (RSVP) -- Version 1 Functional Specification", RFC 2205, September 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.
- [RFC3473] Berger, L., "Generalized Multi-Protocol Label Switching (GMPLS) Signaling Resource ReserVation Protocol-Traffic Engineering (RSVP-TE) Extensions", RFC 3473, January 2003.
- [RFC4090] Pan, P., Swallow, G., and A. Atlas, "Fast Reroute Extensions to RSVP-TE for LSP Tunnels", RFC 4090, May 2005.
- [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.
- [RFC5331] Aggarwal, R., Rekhter, Y., and E. Rosen, "MPLS Upstream Label Assignment and Context-Specific Label Space", RFC 5331, August 2008.
- [RFC5786] Aggarwal, R. and K. Kompella, "Advertising a Router's

Local Addresses in OSPF Traffic Engineering (TE) Extensions", RFC 5786, March 2010.

[P2MP FRR]

Le Roux, J., Aggarwal, R., Vasseur, J., and M. Vigoureux,
"P2MP MPLS-TE Fast Reroute with P2MP Bypass Tunnels",
draft-leroux-mpls-p2mp-te-bypass , March 1997.

11.2. Informative References

[RFC4461] Yasukawa, S., "Signaling Requirements for Point-to-Multipoint Traffic-Engineered MPLS Label Switched Paths (LSPs)", RFC 4461, April 2006.

Authors' Addresses

Huaimo Chen
Huawei Technologies
Boston, MA
USA

Email: huaimo.chen@huawei.com

Zhenbin Li
Huawei Technologies
Huawei Bld., No.156 Beiqing Rd.
Beijing 100095,
China

Email: lizhenbin@huawei.com

Ning So
Tata Communications
2613 Fairbourne Cir.
Plano, TX 75082
USA

Email: ningso01@gmail.com

Autumn Liu
Ericsson
CA
USA

Email: autumn.liu@ericsson.com

Fengman Xu
Verizon
2400 N. Glenville Dr
Richardson, TX 75082
USA

Email: fengman.xu@verizon.com

Mehmet Toy
Comcast
1800 Bishops Gate Blvd.
Mount Laurel, NJ 08054
USA

Email: mehmet_toy@cable.comcast.com

Lu Huang
China Mobile
No.32 Xuanwumen West Street, Xicheng District
Beijing, 100053
China

Email: huanglu@chinamobile.com

Lei Liu
UC Davis
USA

Email: liulei.kddi@gmail.com

Internet Engineering Task Force
Internet-Draft
Intended status: Standards Track
Expires: January 4, 2015

H. Chen, Ed.
Huawei Technologies
R. Torvi, Ed.
Juniper Networks
July 3, 2014

Extensions to RSVP-TE for LSP Ingress Local Protection
draft-ietf-mpls-rsvp-ingress-protection-01.txt

Abstract

This document describes extensions to Resource Reservation Protocol - Traffic Engineering (RSVP-TE) for locally protecting the ingress node of a Traffic Engineered (TE) Label Switched Path (LSP) in a Multi-Protocol Label Switching (MPLS) and Generalized MPLS (GMPLS) network.

Status of this Memo

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

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

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

This Internet-Draft will expire on January 4, 2015.

Copyright Notice

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

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

Table of Contents

1. Co-authors	3
2. Introduction	3
2.1. An Example of Ingress Local Protection	3
2.2. Ingress Local Protection with FRR	4
3. Ingress Failure Detection	4
3.1. Source Detects Failure	4
3.2. Backup and Source Detect Failure	5
3.3. Comparing Different Detection Modes	5
4. Backup Forwarding State	5
4.1. Forwarding State for Backup LSP	6
5. Protocol Extensions	6
5.1. INGRESS_PROTECTION Object	6
5.1.1. Subobject: Backup Ingress IPv4/IPv6 Address	8
5.1.2. Subobject: Ingress IPv4/IPv6 Address	9
5.1.3. Subobject: Traffic Descriptor	9
5.1.4. Subobject: Label-Routes	10
6. Behavior of Ingress Protection	11
6.1. Overview	11
6.1.1. Relay-Message Method	11
6.1.2. Proxy-Ingress Method	11
6.1.3. Comparing Two Methods	12
6.2. Ingress Behavior	13
6.2.1. Relay-Message Method	13
6.2.2. Proxy-Ingress Method	14
6.3. Backup Ingress Behavior	15
6.3.1. Backup Ingress Behavior in Off-path Case	15
6.3.2. Backup Ingress Behavior in On-path Case	17
6.3.3. Failure Detection	18
6.4. Revertive Behavior	19
6.4.1. Revert to Primary Ingress	19
6.4.2. Global Repair by Backup Ingress	19
7. Security Considerations	20
8. IANA Considerations	20
9. Contributors	20
10. Acknowledgement	21
11. References	21
11.1. Normative References	21
11.2. Informative References	22
A. Authors' Addresses	22

1. Co-authors

Ning So, Autumn Liu, Alia Atlas, Yimin Shen, Fengman Xu, Mehmet Toy, Lei Liu

2. Introduction

For MPLS LSPs it is important to have a fast-reroute method for protecting its ingress node as well as transit nodes. This is not covered either in the fast-reroute method defined in [RFC4090] or in the P2MP fast-reroute extensions to fast-reroute in [RFC4875].

An alternate approach to local protection (fast-reroute) is to use global protection and set up a second backup LSP (whether P2MP or P2P) from a backup ingress to the egresses. The main disadvantage of this is that the backup LSP may reserve additional network bandwidth.

This specification defines a simple extension to RSVP-TE for local protection of the ingress node of a P2MP or P2P LSP.

2.1. An Example of Ingress Local Protection

Figure 1 shows an example of using a backup P2MP LSP to locally protect the ingress of a primary P2MP LSP, which is from ingress R1 to three egresses: L1, L2 and L3. The backup LSP is from backup ingress Ra to the next hops R2 and R4 of ingress R1.

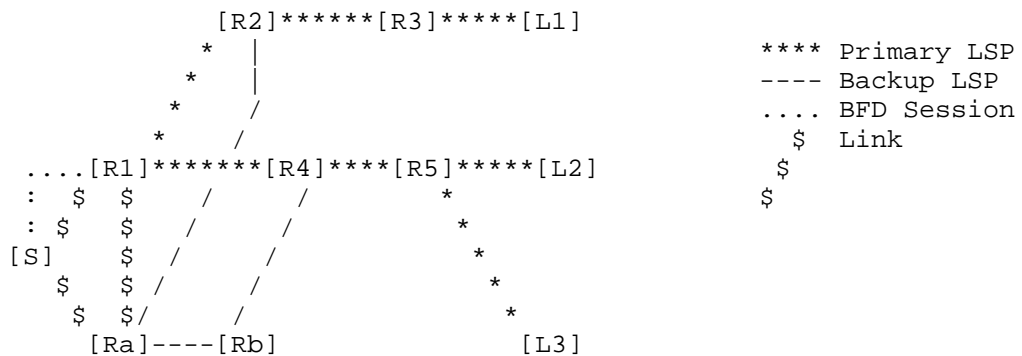


Figure 1: Backup P2MP LSP for Locally Protecting Ingress

In normal operations, source S sends the traffic to primary ingress R1. R1 imports the traffic into the primary LSP to egresses L1, L2 and L3.

When source S detects the failure of R1, it switches the traffic to backup ingress Ra, which imports the traffic from S into the backup LSP to R1's next hops R2 and R4, where the traffic is merged into the primary LSP, and then sent to egresses L1, L2 and L3.

Source S should be able to detect the failure of R1 and switch the traffic within 10s of ms. The exact method by which S does so is out of scope.

Note that the backup ingress must be one logical hop away from the ingress. A logical hop is a direct link or a tunnel such as a GRE tunnel, over which RSVP-TE messages may be exchanged.

2.2. Ingress Local Protection with FRR

Through using the ingress local protection and the FRR, we can locally protect the ingress node, all the links and the intermediate nodes of an LSP. The traffic switchover time is within tens of milliseconds whenever the ingress, any of the links and the intermediate nodes of the LSP fails.

The ingress node of the LSP can be locally protected through using the ingress local protection. All the links and all the intermediate nodes of the LSP can be locally protected through using the FRR.

3. Ingress Failure Detection

Exactly how the failure of the ingress (e.g. R1 in Figure 1) is detected is out of scope for this document. However, it is necessary to discuss different modes for detecting the failure because they determine what must be signaled and what is the required behavior for the traffic source and backup ingress.

3.1. Source Detects Failure

Source Detects Failure or Source-Detect for short means that the source is responsible for fast detecting the failure of the primary ingress of an LSP. The backup ingress is ready to import the traffic from the source into the backup LSP after the backup LSP is up.

In normal operations, the source sends the traffic to the primary ingress. When the source detects the failure of the primary ingress, it switches the traffic to the backup ingress, which delivers the traffic to the next hops of the primary ingress through the backup LSP, where the traffic is merged into the primary LSP.

For a P2P LSP, after the primary ingress fails, the backup ingress

must use a method to reliably detect the failure of the primary ingress before the PATH message for the LSP expires at the next hop of the primary ingress. After reliably detecting the failure, the backup ingress sends/refreshes the PATH message to the next hop through the backup LSP as needed.

After the primary ingress fails, it will not be reachable after routing convergence. Thus checking whether the primary ingress (address) is reachable is a possible method.

3.2. Backup and Source Detect Failure

Backup and Source Detect Failure or Backup-Source-Detect for short means that both the backup ingress and the source are concurrently responsible for fast detecting the failures of the primary ingress.

In normal operations, the source sends the traffic to the primary ingress. It switches the traffic to the backup ingress when it detects the failure of the primary ingress.

The backup ingress does not import any traffic from the source into the backup LSP in normal operations. When it detects the failure of the primary ingress, it imports the traffic from the source into the backup LSP to the next hops of the primary ingress, where the traffic is merged into the primary LSP.

Note that the source may locally distinguish between the failure of the primary ingress and that of the link between the source and the primary ingress. When the source detects the failure of the link, it may continue to send the traffic to the primary ingress via another link between the source and the primary ingress if there is one.

3.3. Comparing Different Detection Modes

The source-detect is preferred. It is simpler than the backup-source-detect, which needs both the source and the backup ingress detect the ingress failure quickly.

4. Backup Forwarding State

Before the primary ingress fails, the backup ingress is responsible for creating the necessary backup LSPs. These LSPs might be multiple bypass P2P LSPs that avoid the ingress. Alternately, the backup ingress could choose to use a single backup P2MP LSP as a bypass or detour to protect the primary ingress of a primary P2MP LSP.

The backup ingress may be off-path or on-path of an LSP. When a

backup ingress is not any node of the LSP, we call the backup ingress is off-path. When a backup ingress is a next-hop of the primary ingress of the LSP, we call it is on-path. If the backup ingress is on-path, the primary forwarding state associated with the primary LSP SHOULD be clearly separated from the backup LSP(s) state.

4.1. Forwarding State for Backup LSP

A forwarding entry for a backup LSP is created on the backup ingress after the LSP is set up. Depending on the failure-detection mode (e.g., source-detect), it may be used to forward received traffic or simply be inactive (e.g., backup-source-detect) until required. In either case, when the primary ingress fails, this entry is used to import the traffic into the backup LSP to the next hops of the primary ingress, where the traffic is merged into the primary LSP.

The forwarding entry for a backup LSP is a local implementation issue. In one device, it may have an inactive flag. This inactive forwarding entry is not used to forward any traffic normally. When the primary ingress fails, it is changed to active, and thus the traffic from the source is imported into the backup LSP.

5. Protocol Extensions

A new object `INGRESS_PROTECTION` is defined for signaling ingress local protection. It is backward compatible.

5.1. `INGRESS_PROTECTION` Object

The `INGRESS_PROTECTION` object with the `FAST_REROUTE` object in a `PATH` message is used to control the backup for protecting the primary ingress of a primary LSP. The primary ingress MUST insert this object into the `PATH` message to be sent to the backup ingress for protecting the primary ingress. It has the following format:

```

Class-Num = TBD          C-Type = TBD
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
+-----+-----+-----+-----+-----+-----+-----+-----+
|          Length (bytes)          |          Class-Num          |          C-Type          |
+-----+-----+-----+-----+-----+-----+-----+-----+
|          Secondary LSP ID          |          Flags          |          Options          |
+-----+-----+-----+-----+-----+-----+-----+-----+
~                                     (Subobjects)                                     ~
+-----+-----+-----+-----+-----+-----+-----+-----+

```

Flags

- 0x01 Ingress local protection available
- 0x02 Ingress local protection in use
- 0x04 Bandwidth protection

Options

- 0x01 Revert to Ingress
- 0x02 Ingress-Proxy/Relay-Message
- 0x04 P2MP Backup

The Secondary LSP ID in the object is an LSP ID that the primary ingress has allocated for a protected LSP tunnel. The backup ingress will use this LSP ID to set up a new LSP from the backup ingress to the destinations of the protected LSP tunnel. This allows the new LSP to share resources with the old one.

The flags are used to communicate status information from the backup ingress to the primary ingress.

- o Ingress local protection available: The backup ingress sets this flag after backup LSPs are up and ready for locally protecting the primary ingress. The backup ingress sends this to the primary ingress to indicate that the primary ingress is locally protected.
- o Ingress local protection in use: The backup ingress sets this flag when it detects a failure in the primary ingress. The backup ingress keeps it and does not send it to the primary ingress since the primary ingress is down.
- o Bandwidth protection: The backup ingress sets this flag if the backup LSPs guarantee to provide desired bandwidth for the protected LSP against the primary ingress failure.

The options are used by the primary ingress to specify the desired behavior to the backup ingress and next-hops.

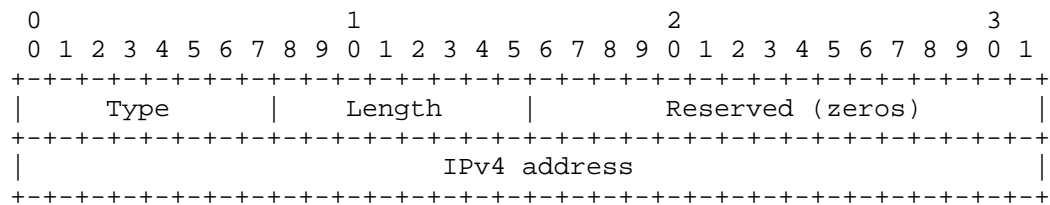
- o Revert to Ingress: The primary ingress sets this option indicating that the traffic for the primary LSP successfully re-signaled will be switched back to the primary ingress from the backup ingress when the primary ingress is restored.
- o Ingress-Proxy/Relay-Message: This option is set to one indicating that Ingress-Proxy method is used. It is set to zero indicating that Relay-Message method is used.

- o P2MP Backup: This option is set to ask for the backup ingress to use P2MP backup LSP to protect the primary ingress. Note that one spare bit of the flags in the FAST-REROUTE object can be used to indicate whether P2MP or P2P backup LSP is desired for protecting an ingress and intermediate node.

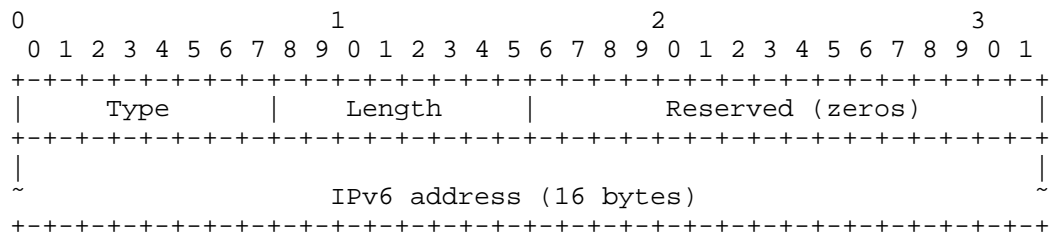
The INGRESS_PROTECTION object may contain some of the sub objects described below.

5.1.1.1. Subobject: Backup Ingress IPv4/IPv6 Address

When the primary ingress of a protected LSP sends a PATH message with an INGRESS_PROTECTION object to the backup ingress, the object may have a Backup Ingress IPv4/IPv6 Address sub object containing an IPv4/IPv6 address belonging to the backup ingress. The formats of the sub object for Backup Ingress IPv4/IPv6 Address is given below:



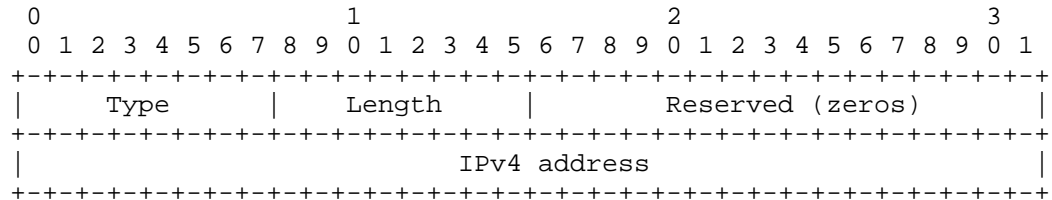
Type: TBD-1 Backup Ingress IPv4 Address
 Length: Total length of the subobject in bytes, including the Type and Length fields. The Length is always 8.
 Reserved: Reserved two bytes are set to zeros.
 IPv4 address: A 32-bit unicast, host address.



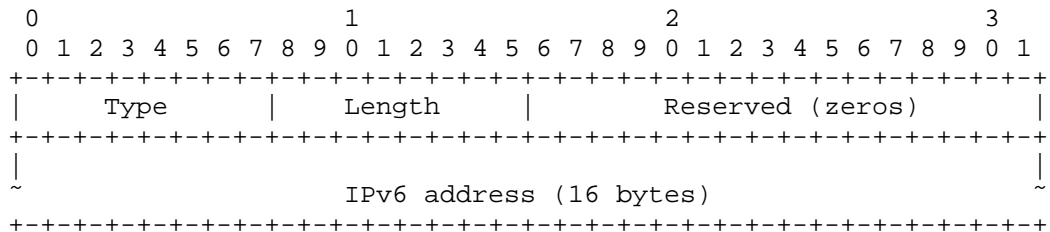
Type: TBD-2 Backup Ingress IPv6 Address
 Length: Total length of the subobject in bytes, including the Type and Length fields. The Length is always 20.
 Reserved: Reserved two bytes are set to zeros.
 IPv6 address: A 128-bit unicast, host address.

5.1.2. Subobject: Ingress IPv4/IPv6 Address

The INGRESS_PROTECTION object may have an Ingress IPv4/IPv6 Address sub object containing an IPv4/IPv6 address belonging to the primary ingress. The sub object has the following format:



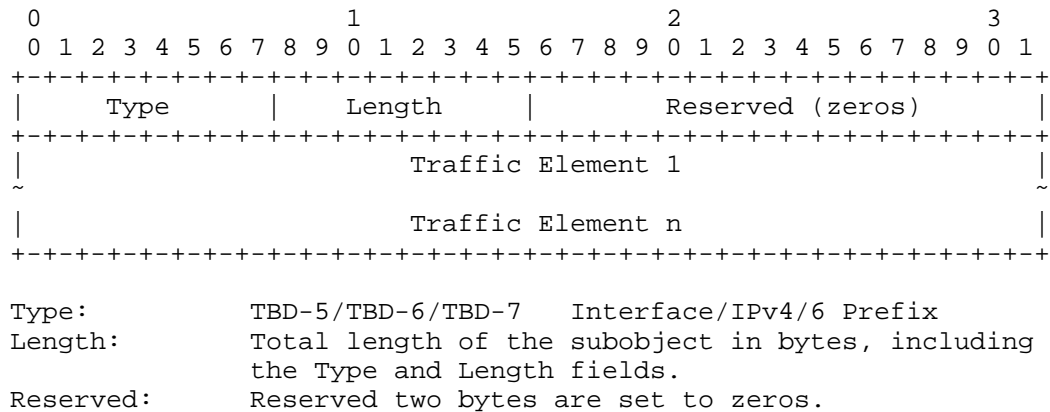
Type: TBD-3 Ingress IPv4 Address
 Length: Total length of the subobject in bytes, including the Type and Length fields. The Length is always 8.
 Reserved: Reserved two bytes are set to zeros.
 IPv4 address: A 32-bit unicast, host address.



Type: TBD-4 Backup Ingress IPv6 Address
 Length: Total length of the subobject in bytes, including the Type and Length fields. The Length is always 20.
 Reserved: Reserved two bytes are set to zeros.
 IPv6 address: A 128-bit unicast, host address.

5.1.3. Subobject: Traffic Descriptor

The INGRESS_PROTECTION object may have a Traffic Descriptor sub object describing the traffic to be mapped to the backup LSP on the backup ingress for locally protecting the primary ingress. The sub object has the following format:

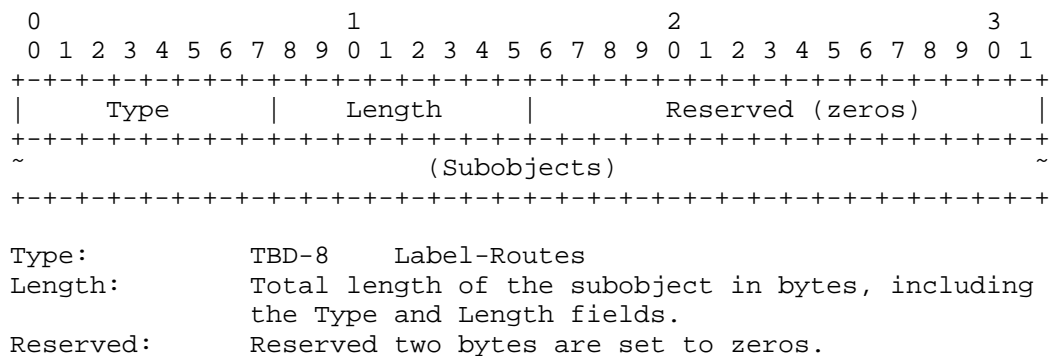


The Traffic Descriptor sub object may contain multiple Traffic Elements of same type as follows.

- o Interface Traffic (Type TBD-5): Each of the Traffic Elements is a 32 bit index of an interface, from which the traffic is imported into the backup LSP.
- o IPv4/6 Prefix Traffic (Type TBD-6/TBD-7): Each of the Traffic Elements is an IPv4/6 prefix, containing an 8-bit prefix length followed by an IPv4/6 address prefix, whose length, in bits, was specified by the prefix length, padded to a byte boundary.

5.1.4. Subobject: Label-Routes

The INGRESS_PROTECTION object in a PATH message from the primary ingress to the backup ingress will have a Label-Routes sub object containing the labels and routes that the next hops of the ingress use. The sub object has the following format:



The Subobjects in the Label-Routes are copied from the Subobjects in the RECORD_ROUTE objects contained in the RESV messages that the primary ingress receives from its next hops for the protected LSP. They MUST contain the first hops of the LSP, each of which is paired with its label.

6. Behavior of Ingress Protection

6.1. Overview

There are four parts of ingress protection: 1) setting up the necessary backup LSP forwarding state; 2) identifying the failure and providing the fast repair (as discussed in Sections 2 and 3); 3) maintaining the RSVP-TE control plane state until a global repair can be done; and 4) performing the global repair(see Section 5.5).

There are two different proposed signaling approaches to obtain ingress protection. They both use the same new INGRESS-PROTECTION object. The object is sent in both PATH and RESV messages.

6.1.1. Relay-Message Method

The primary ingress relays the information for ingress protection of an LSP to the backup ingress via PATH messages. Once the LSP is created, the ingress of the LSP sends the backup ingress a PATH message with an INGRESS-PROTECTION object with Label-Routes subobject, which is populated with the next-hops and labels. This provides sufficient information for the backup ingress to create the appropriate forwarding state and backup LSP(s).

The ingress also sends the backup ingress all the other PATH messages for the LSP with an empty INGRESS-PROTECTION object. Thus, the backup ingress has access to all the PATH messages needed for modification to be sent to refresh control-plane state after a failure.

The advantages of this method include: 1) the primary LSP is independent of the backup ingress; 2) simple; 3) less configuration; and 4) less control traffic.

6.1.2. Proxy-Ingress Method

Conceptually, a proxy ingress is created that starts the RSVP signaling. The explicit path of the LSP goes from the proxy ingress to the backup ingress and then to the real ingress. The behavior and signaling for the proxy ingress is done by the real ingress; the use of a proxy ingress address avoids problems with loop detection.

```

          [ traffic source ]          *** Primary LSP
          $                      $    --- Backup LSP
          $                      $    $$  Link
          $                      $
[ proxy ingress ] [ backup ]
[ & ingress      ]      |
      *              |
      *****[ MP ]----|

```

Figure 2: Example Protected LSP with Proxy Ingress Node

The backup ingress must know the merge points or next-hops and their associated labels. This is accomplished by having the RSVP PATH and RESV messages go through the backup ingress, although the forwarding path need not go through the backup ingress. If the backup ingress fails, the ingress simply removes the INGRESS-PROTECTION object and forwards the PATH messages to the LSP's next-hop(s). If the ingress has its LSP configured for ingress protection, then the ingress can add the backup ingress and itself to the ERO and start forwarding the PATH messages to the backup ingress.

Slightly different behavior can apply for the on-path and off-path cases. In the on-path case, the backup ingress is a next hop node after the ingress for the LSP. In the off-path, the backup ingress is not any next-hop node after the ingress for all associated sub-LSPs.

The key advantage of this approach is that it minimizes the special handling code requires. Because the backup ingress is on the signaling path, it can receive various notifications. It easily has access to all the PATH messages needed for modification to be sent to refresh control-plane state after a failure.

6.1.3. Comparing Two Methods

Method	Primary LSP Depends on Backup Ingress	Simple	Config Proxy- Ingress- ID	PATH Msg from Backup to primary RESV Msg from Primary to backup	Reuse Some of Existing Functions
Relay- Message	No	Yes	No	No	Yes-
Proxy- Ingress	Yes	Yes-	Yes	Yes	Yes

6.2. Ingress Behavior

The primary ingress must be configured with two or three pieces of information for ingress protection.

- o Backup Ingress Address: The primary ingress must know an IP address for it to be included in the INGRESS-PROTECTION object.
- o Proxy-Ingress-Id (only needed for Proxy-Ingress Method): The Proxy-Ingress-Id is only used in the Record Route Object for recording the proxy-ingress. If no proxy-ingress-id is specified, then a local interface address that will not otherwise be included in the Record Route Object can be used. A similar technique is used in [RFC4090 Sec 6.1.1].
- o Application Traffic Identifier: The primary ingress and backup ingress must both know what application traffic should be directed into the LSP. If a list of prefixes in the Traffic Descriptor sub-object will not suffice, then a commonly understood Application Traffic Identifier can be sent between the primary ingress and backup ingress. The exact meaning of the identifier should be configured similarly at both the primary ingress and backup ingress. The Application Traffic Identifier is understood within the unique context of the primary ingress and backup ingress.

With this additional information, the primary ingress can create and signal the necessary RSVP extensions to support ingress protection.

6.2.1. Relay-Message Method

To protect the ingress of an LSP, the ingress does the following after the LSP is up.

1. Select a PATH message.
2. If the backup ingress is off-path, then send the backup ingress a PATH message with the content from the selected PATH message and an INGRESS-PROTECTION object; else (the backup ingress is a next hop, i.e., on-path case) add an INGRESS-PROTECTION object into the existing PATH message to the backup ingress (i.e., the next hop). The INGRESS-PROTECTION object contains the Traffic-Descriptor sub-object, the Backup Ingress Address sub-object and the Label-Routes sub-object. The flags is set to indicate whether a Backup P2MP LSP is desired. If not yet allocated, allocate a second LSP-ID to be used in the INGRESS-PROTECTION object. The Label-Routes sub-object contains the next-hops of the ingress and their labels.

3. For each of the other PATH messages, if the node to which the message is sent is not the backup ingress, then send the backup ingress a PATH message with the content copied from the message to the node and an empty INGRESS-PROTECTION object; else send the node the message with an empty INGRESS-PROTECTION object.

6.2.2. Proxy-Ingress Method

The primary ingress is responsible for starting the RSVP signaling for the proxy-ingress node. To do this, the following is done for the RSVP PATH message.

1. Compute the EROs for the LSP as normal for the ingress.
2. If the selected backup ingress node is not the first node on the path (for all sub-LSPs), then insert at the beginning of the ERO first the backup ingress node and then the ingress node.
3. In the PATH RRO, instead of recording the ingress node's address, replace it with the Proxy-Ingress-Id.
4. Leave the HOP object populated as usual with information for the ingress-node.
5. Add the INGRESS-PROTECTION object to the PATH message. Allocate a second LSP-ID to be used in the INGRESS-PROTECTION object. Include the Backup Ingress Address (IPv4 or IPv6) sub-object and the Traffic-Descriptor sub-object. Set or clear the flag indicating that a Backup P2MP LSP is desired.
6. Optionally, add the FAST-REROUTE object [RFC4090] to the Path message. Indicate whether one-to-one backup is desired. Indicate whether facility backup is desired.
7. The RSVP PATH message is sent to the backup node as normal.

If the ingress detects that it can't communicate with the backup ingress, then the ingress should instead send the PATH message to the next-hop indicated in the ERO computed in step 1. Once the ingress detects that it can communicate with the backup ingress, the ingress SHOULD follow the steps 1-7 to obtain ingress failure protection.

When the ingress node receives an RSVP PATH message with an INGRESS-PROTECTION object and the object specifies that node as the ingress node and the PHOP as the backup ingress node, the ingress node SHOULD remove the INGRESS-PROTECTION object from the PATH message before sending it out. Additionally, the ingress node must store that it will install ingress forwarding state for the LSP rather than

midpoint forwarding.

When an RSVP RESV message is received by the ingress, it uses the NHOP to determine whether the message is received from the backup ingress or from a different node. The stored associated PATH message contains an INGRESS-PROTECTION object that identifies the backup ingress node. If the RESV message is not from the backup node, then ingress forwarding state should be set up, and the INGRESS-PROTECTION object MUST be added to the RESV before it is sent to the NHOP, which should be the backup node. If the RESV message is from the backup node, then the LSP should be considered available for use.

If the backup ingress node is on the forwarding path, then a RESV is received with an INGRESS-PROTECTION object and an NHOP that matches the backup ingress. In this case, the ingress node's address will not appear after the backup ingress in the RRO. The ingress node should set up ingress forwarding state, just as is done if the LSP weren't ingress-node protected.

6.3. Backup Ingress Behavior

An LER determines that the ingress local protection is requested for an LSP if the INGRESS_PROTECTION object is included in the PATH message it receives for the LSP. The LER can further determine that it is the backup ingress if one of its addresses is in the Backup Ingress Address sub-object of the INGRESS-PROTECTION object. The LER as the backup ingress will assume full responsibility of the ingress after the primary ingress fails. In addition, the LER determines that it is off-path if it is not a next hop of the primary ingress.

6.3.1. Backup Ingress Behavior in Off-path Case

The backup ingress considers itself as a PLR and the primary ingress as its next hop and provides a local protection for the primary ingress. It behaves very similarly to a PLR providing fast-reroute where the primary ingress is considered as the failure-point to protect. Where not otherwise specified, the behavior given in [RFC4090] for a PLR should apply.

The backup ingress SHOULD follow the control-options specified in the INGRESS-PROTECTION object and the flags and specifications in the FAST-REROUTE object. This applies to providing a P2MP backup if the "P2MP backup" is set, a one-to-one backup if "one-to-one desired" is set, facility backup if the "facility backup desired" is set, and backup paths that support the desired bandwidth, and administrative-colors that are requested.

If multiple INGRESS-PROTECTION objects have been received via

multiple PATH messages for the same LSP, then the most recent one that specified a Traffic-Descriptor sub-object MUST be the one used.

The backup ingress creates the appropriate forwarding state for the backup LSP tunnel(s) to the merge point(s).

When the backup ingress sends a RESV message to the primary ingress, it should add an INGRESS-PROTECTION object into the message. It SHOULD set or clear the flags in the object to report "Ingress local protection available", "Ingress local protection in use", and "bandwidth protection".

If the backup ingress doesn't have a backup LSP tunnel to all the merge points, it SHOULD clear "Ingress local protection available". [Editor Note: It is possible to indicate the number or which are unprotected via a sub-object if desired.]

When the primary ingress fails, the backup ingress redirects the traffic from a source into the backup P2P LSPs or the backup P2MP LSP transmitting the traffic to the next hops of the primary ingress, where the traffic is merged into the protected LSP.

In this case, the backup ingress keeps the PATH message with the INGRESS_PROTECTION object received from the primary ingress and the RESV message with the INGRESS_PROTECTION object to be sent to the primary ingress. The backup ingress sets the "local protection in use" flag in the RESV message, indicating that the backup ingress is actively redirecting the traffic into the backup P2P LSPs or the backup P2MP LSP for locally protecting the primary ingress failure.

Note that the RESV message with this piece of information will not be sent to the primary ingress because the primary ingress has failed.

If the backup ingress has not received any PATH message from the primary ingress for an extended period of time (e.g., a cleanup timeout interval) and a confirmed primary ingress failure did not occur, then the standard RSVP soft-state removal SHOULD occur. The backup ingress SHALL remove the state for the PATH message from the primary ingress, and tear down the one-to-one backup LSPs for protecting the primary ingress if one-to-one backup is used or unbind the facility backup LSPs if facility backup is used.

When the backup ingress receives a PATH message from the primary ingress for locally protecting the primary ingress of a protected LSP, it checks to see if any critical information has been changed. If the next hops of the primary ingress are changed, the backup ingress SHALL update its backup LSP(s).

6.3.1.1. Relay-Message Method

When the backup ingress receives a PATH message with the INGRESS-PROTECTION object, it examines the object to learn what traffic associated with the LSP. It determines the next-hops to be merged to by examining the Label-Routes sub-object in the object. If the Traffic-Descriptor sub-object isn't included, this object is considered "empty".

The backup ingress stores the PATH message received from the primary ingress, but does NOT forward it.

The backup ingress MUST respond with a RESV to the PATH message received from the primary ingress. If the INGRESS-PROTECTION object is not "empty", the backup ingress SHALL send the RESV message with the state indicating protection is available after the backup LSP(s) are successfully established.

6.3.1.2. Proxy-Ingress Method

The backup ingress determines the next-hops to be merged to by collecting the set of the pair of (IPv4/IPv6 sub-object, Label sub-object) from the Record Route Object of each RESV that are closest to the top and not the Ingress router; this should be the second to the top pair. If a Label-Routes sub-object is included in the INGRESS-PROTECTION object, the included IPv4/IPv6 sub-objects are used to filter the set down to the specific next-hops where protection is desired. A RESV message must have been received before the Backup Ingress can create or select the appropriate backup LSP.

When the backup ingress receives a PATH message with the INGRESS-PROTECTION object, the backup ingress examines the object to learn what traffic associated with the LSP. The backup ingress forwards the PATH message to the ingress node with the normal RSVP changes.

When the backup ingress receives a RESV message with the INGRESS-PROTECTION object, the backup ingress records an IMPLICIT-NULL label in the RRO. Then the backup ingress forwards the RESV message to the ingress node, which is acting for the proxy ingress.

6.3.2. Backup Ingress Behavior in On-path Case

An LER as the backup ingress determines that it is on-path if one of its addresses is a next hop of the primary ingress and the primary ingress is not its next hop via checking the PATH message with the INGRESS_PROTECTION object received from the primary ingress. The LER on-path sends the corresponding PATH messages without any INGRESS_PROTECTION object to its next hops. It creates a number of

backup P2P LSPs or a backup P2MP LSP from itself to the other next hops (i.e., the next hops other than the backup ingress) of the primary ingress. The other next hops are from the Label-Routes sub object.

It also creates a forwarding entry, which sends/multicasts the traffic from the source to the next hops of the backup ingress along the protected LSP when the primary ingress fails. The traffic is described by the Traffic-Descriptor.

After the forwarding entry is created, all the backup P2P LSPs or the backup P2MP LSP is up and associated with the protected LSP, the backup ingress sends the primary ingress the RESV message with the INGRESS_PROTECTION object containing the state of the local protection such as "local protection available" flag set to one, which indicates that the primary ingress is locally protected.

When the primary ingress fails, the backup ingress sends/multicasts the traffic from the source to its next hops along the protected LSP and imports the traffic into each of the backup P2P LSPs or the backup P2MP LSP transmitting the traffic to the other next hops of the primary ingress, where the traffic is merged into protected LSP.

During the local repair, the backup ingress continues to send the PATH messages to its next hops as before, keeps the PATH message with the INGRESS_PROTECTION object received from the primary ingress and the RESV message with the INGRESS_PROTECTION object to be sent to the primary ingress. It sets the "local protection in use" flag in the RESV message.

6.3.3. Failure Detection

As described in [RFC4090], it is necessary to refresh the PATH messages via the backup LSP(s). The Backup Ingress MUST wait to refresh the backup PATH messages until it can accurately detect that the ingress node has failed. An example of such an accurate detection would be that the IGP has no bi-directional links to the ingress node and the last change was long enough in the past that changes should have been received (i.e., an IGP network convergence time or approximately 2-3 seconds) or a BFD session to the primary ingress' loopback address has failed and stayed failed after the network has reconverged.

As described in [RFC4090 Section 6.4.3], the backup ingress, acting as PLR, SHOULD modify - including removing any INGRESS-PROTECTION and FAST-REROUTE objects - and send any saved PATH messages associated with the primary LSP.

6.4. Revertive Behavior

Upon a failure event in the (primary) ingress of a protected LSP, the protected LSP is locally repaired by the backup ingress. There are a couple of basic strategies for restoring the LSP to a full working path.

- Revert to Primary Ingress: When the primary ingress is restored, it re-signals each of the LSPs that start from the primary ingress. The traffic for every LSP successfully re-signaled is switched back to the primary ingress from the backup ingress.
- Global Repair by Backup Ingress: After determining that the primary ingress of an LSP has failed, the backup ingress computes a new optimal path, signals a new LSP along the new path, and switches the traffic to the new LSP.

6.4.1. Revert to Primary Ingress

If "Revert to Primary Ingress" is desired for a protected LSP, the (primary) ingress of the LSP re-signals the LSP that starts from the primary ingress after the primary ingress restores. When the LSP is re-signaled successfully, the traffic is switched back to the primary ingress from the backup ingress and redirected into the LSP starting from the primary ingress.

If the ingress can resignal the PATH messages for the LSP, then the ingress can specify the "Revert to Ingress" control-option in the INGRESS-PROTECTION object. Doing so may cause a duplication of traffic while the Ingress starts sending traffic again before the Backup Ingress stops; the alternative is to drop traffic for a short period of time.

Additionally, the Backup Ingress can set the "Revert To Ingress" control-option as a request for the Ingress to take over.

6.4.2. Global Repair by Backup Ingress

When the backup ingress has determined that the primary ingress of the protected LSP has failed (e.g., via the IGP), it can compute a new path and signal a new LSP along the new path so that it no longer relies upon local repair. To do this, the backup ingress uses the same tunnel sender address in the Sender Template Object and uses the previously allocated second LSP-ID in the INGRESS-PROTECTION object of the PATH message as the LSP-ID of the new LSP. This allows the new LSP to share resources with the old LSP. In addition, if the Ingress recovers, the Backup Ingress SHOULD send it RESVs with the INGRESS-PROTECTION object where either the "Force to Backup" or

"Revert to Ingress" is specified. The Secondary LSP ID should be the unused LSP ID - while the LSP ID signaled in the RESV will be that currently active. The Ingress can learn from the RESVs what to signal. Even if the Ingress does not take over, the RESVs notify it that the particular LSP IDs are in use. The Backup Ingress can reoptimize the new LSP as necessary until the Ingress recovers. Alternately, the Backup Ingress can create a new LSP with no bandwidth reservation that duplicates the path(s) of the protected LSP, move traffic to the new LSP, delete the protected LSP, and then resignal the new LSP with bandwidth.

7. Security Considerations

In principle this document does not introduce new security issues. The security considerations pertaining to RFC 4090, RFC 4875 and other RSVP protocols remain relevant.

8. IANA Considerations

TBD

9. Contributors

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

Zhenbin Li
Huawei Technologies
2330 Central Expressway
Santa Clara, CA 95050
USA
Email: zhenbin.li@huawei.com

Boris Zhang
Telus Communications
200 Consilium Pl Floor 15
Toronto, ON M1H 3J3
Canada
Email: Boris.Zhang@telus.com

Markus Jork
Juniper Networks
10 Technology Park Drive
Westford, MA 01886
USA
Email: mjork@juniper.net

10. Acknowledgement

The authors would like to thank Nobo Akiya, Rahul Aggarwal, Eric Osborne, Ross Callon, Loa Andersson, Michael Yue, Olufemi Komolafe, Rob Rennison, Neil Harrison, Kannan Sampath, and Ronhazli Adam for their valuable comments and suggestions on this draft.

11. References

11.1. Normative References

- [RFC1700] Reynolds, J. and J. Postel, "Assigned Numbers", RFC 1700, October 1994.
- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC3692] Narten, T., "Assigning Experimental and Testing Numbers Considered Useful", BCP 82, RFC 3692, January 2004.

- [RFC2205] Braden, B., Zhang, L., Berson, S., Herzog, S., and S. Jamin, "Resource ReSerVation Protocol (RSVP) -- Version 1 Functional Specification", RFC 2205, September 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.
- [RFC3473] Berger, L., "Generalized Multi-Protocol Label Switching (GMPLS) Signaling Resource ReserVation Protocol-Traffic Engineering (RSVP-TE) Extensions", RFC 3473, January 2003.
- [RFC4090] Pan, P., Swallow, G., and A. Atlas, "Fast Reroute Extensions to RSVP-TE for LSP Tunnels", RFC 4090, May 2005.
- [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.
- [P2MP-FRR] Le Roux, J., Aggarwal, R., Vasseur, J., and M. Vigoureux, "P2MP MPLS-TE Fast Reroute with P2MP Bypass Tunnels", draft-leroux-mpls-p2mp-te-bypass , March 1997.

11.2. Informative References

- [RFC2702] Awduche, D., Malcolm, J., Agogbua, J., O'Dell, M., and J. McManus, "Requirements for Traffic Engineering Over MPLS", RFC 2702, September 1999.
- [RFC3032] Rosen, E., Tappan, D., Fedorkow, G., Rekhter, Y., Farinacci, D., Li, T., and A. Conta, "MPLS Label Stack Encoding", RFC 3032, January 2001.

Appendix A. Authors' Addresses

Huaimo Chen
Huawei Technologies
Boston, MA
USA
Email: huaimo.chen@huawei.com

Ning So
Tata Communications
2613 Fairbourne Cir.
Plano, TX 75082
USA
Email: ningso01@gmail.com

Autumn Liu
Ericsson
300 Holger Way
San Jose, CA 95134
USA
Email: autumn.liu@ericsson.com

Raveendra Torvi
Juniper Networks
10 Technology Park Drive
Westford, MA 01886
USA
Email: rtorvi@juniper.net

Alia Atlas
Juniper Networks
10 Technology Park Drive
Westford, MA 01886
USA
Email: akatlas@juniper.net

Yimin Shen
Juniper Networks
10 Technology Park Drive
Westford, MA 01886
USA
Email: yshen@juniper.net

Fengman Xu
Verizon
2400 N. Glenville Dr
Richardson, TX 75082
USA
Email: fengman.xu@verizon.com

Mehmet Toy
Comcast
1800 Bishops Gate Blvd.
Mount Laurel, NJ 08054
USA
Email: mehmet_toy@cable.comcast.com

Lei Liu
UC Davis
USA
Email: liulei.kddi@gmail.com

MPLS WG
Internet-Draft
Intended status: Standards Track
Expires: January 5, 2015

K. Kompella
R. Balaji
Juniper Networks, Inc.
July 4, 2014

Label Distribution Using ARP
draft-kompella-mpls-larp-01

Abstract

This document describes extensions to the Address Resolution Protocol to distribute MPLS labels for IPv4 and IPv6 host addresses. Distribution of labels via ARP enables simple plug-and-play operation of MPLS, which is a key goal of the MPLS Fabric architecture.

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].

The term "server" will be used in this document to refer to an ARP/L-ARP server; the term "host" will be used to refer to a compute server or other device acting as an ARP/L-ARP client.

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 5, 2015.

Copyright Notice

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

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

Table of Contents

1. Introduction	2
1.1. Approach	3
2. Overview of Ethernet ARP	3
3. L-ARP Protocol Operation	4
3.1. Basic Operation	4
3.2. Asynchronous operation	5
3.3. Client-Server Synchronization	5
3.4. Applicability	6
3.5. Backward Compatibility	6
4. For Future Study	6
5. L-ARP Message Format	7
6. Security Considerations	9
7. IANA Considerations	9
8. Acknowledgments	9
9. Normative References	9
Authors' Addresses	10

1. Introduction

This document describes extensions to the Address Resolution Protocol (ARP) [RFC0826] to advertise label bindings for IP host addresses. While there are well-established protocols, such as LDP, RSVP and BGP, that provide robust mechanisms for label distribution, these protocols tend to be relatively complex, and often require detailed configuration for proper operation. There are situations where a simpler protocol may be more suitable from an operational standpoint. An example is the case where an MPLS Fabric is the underlay technology in a Data Centre; here, MPLS tunnels originate from host machines. The host thus needs a mechanism to acquire label bindings to participate in the MPLS Fabric, but in a simple, plug-and-play manner. Existing signaling/routing protocols do not always meet this need. Labeled ARP (L-ARP) is a proposal to fill that gap.

[TODO-MPLS-FABRIC] describes the motivation for using MPLS as the fabric technology.

1.1. Approach

ARP is a nearly ubiquitous protocol; every device with an Ethernet interface, from hand-helds to hosts, have an implementation of ARP. ARP is plug-and-play; ARP clients do not need configuration to use ARP. That suggests that ARP may be a good fit for devices that want to source and sink MPLS tunnels, but do so in a zero-config, plug-and-play manner, with minimal impact to their code.

The approach taken here is to create a minor variant of the ARP protocol, labeled ARP (L-ARP), which is distinguished by a new hardware type, MPLS-over-Ethernet. Regular (Ethernet) ARP (E-ARP) and L-ARP can coexist; a device, as an ARP client, can choose to send out an E-ARP or an L-ARP request, depending on whether it needs Ethernet or MPLS connectivity. Another device may choose to function as an E-ARP server and/or an L-ARP server, depending on its ability to provide an IP-to-Ethernet and/or IP-to-MPLS mapping.

2. Overview of Ethernet ARP

In the most straightforward mode of operation [RFC0826], ARP queries are sent to resolve "directly connected" IP addresses. The ARP query is broadcast, with the Target Protocol Address field (see Section 5 for a description of the fields in an ARP message) carrying the IP address of another node in the same subnet. All the nodes in the LAN receive this ARP query. All the nodes, except the node that owns the IP address, ignore the ARP query. The IP address owner learns the MAC address of the sender from the Source Hardware Address field in the ARP request, and unicasts an ARP reply to the sender. The ARP reply carries the replying node's MAC address in the Source Hardware Address field, thus enabling two-way communication between the two nodes.

A variation of this scheme, known as "proxy ARP" [RFC2002], allows a node to respond to an ARP request with its own MAC address, even when the responding node does not own the requested IP address. Generally, the proxy ARP response is generated by routers to attract traffic for prefixes they can forward packets to. This scheme requires the host to send ARP queries for the IP address the host is trying to reach, rather than the IP address of the router. When there is more than one router connected to a network, proxy ARP enables a host to automatically select an exit router without running any routing protocol to determine IP reachability. Unlike regular ARP, a proxy ARP request can elicit multiple responses, e.g., when more than one router has connectivity to the address being resolved. The sender must be prepared to select one of the responding routers.

Yet another variation of the ARP protocol, called 'Gratuitous ARP' [RFC2002], allows a node to update the ARP cache of other nodes in an unsolicited fashion. Gratuitous ARP is sent as either an ARP request or an ARP reply. In either case, the Source Protocol Address and Target Protocol Address contain the sender's address, and the Source Hardware Address is set to the sender's hardware address. In case of a gratuitous ARP reply, the Target Hardware Address is also set to the sender's address.

3. L-ARP Protocol Operation

The L-ARP protocol builds on the proxy ARP model, and also leverages gratuitous ARP model for asynchronous updates.

In this memo, we will refer to L-ARP clients (that make L-ARP requests) and L-ARP servers (that send L-ARP responses). In Figure 1, H1, H2 and H3 are L-ARP clients, and T1, T2 and T3 are L-ARP servers. T is a member of the MPLS Fabric that may not be an L-ARP server. Within the MPLS Fabric, the usual MPLS protocols (IGP, LDP, RSVP-TE) are run. Say H1, H2 and H3 want to establish MPLS tunnels to each other (for example, they are using BGP MPLS VPNs as the overlay virtual network technology). H1 might also want to talk to a member of the MPLS Fabric, say T.

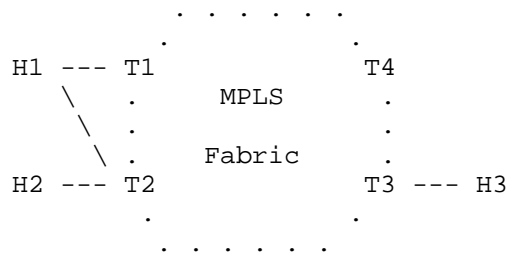


Figure 1

3.1. Basic Operation

A node (say H1) that needs an MPLS tunnel to a destination (say H3) broadcasts over all its interfaces an L-ARP query with the Target Protocol Address set to H3. A node that has reachability to H3 (such as T1 or T2) sends an L-ARP reply with the Source Hardware Address set to a locally-allocated MPLS label plus its Ethernet MAC address. After receiving one or more L-ARP replies, H1 can select either T1 or T2 to send MPLS packets that are destined to H3. As described later, the L-ARP response may contain certain parameters that enable the client to make an informed choice of the routers.

As with standard ARP, the validity of the MPLS label obtained using L-ARP is time-bound. The client should periodically resend its L-ARP requests to obtain the latest information, and time out entries in its ARP cache if such an update is not forthcoming. Once an L-ARP server has advertised a label binding, it MUST NOT change the binding until expiry of the binding's validity time.

The mechanism defined here is simplistic; see Section 4.

3.2. Asynchronous operation

The preceding sections described a request-response based model. In some cases, the L-ARP server may want to asynchronously update its clients. L-ARP uses the gratuitous ARP model [RFC2002] to "push" such changes.

In a pure "push" model, a device may send out updates for all prefixes it knows about. This naive approach will not scale well. This memo specifies a mode of operation that is somewhere between "push" and "pull" model. An L-ARP server does not advertise any binding for a prefix until at least one L-ARP client expresses interest in that prefix (by initiating an L-ARP query). As long as the server has at least one interested client for a prefix, the server sends unsolicited (aka gratuitous, though the term is less appropriate in this context) L-ARP replies when a prefix's reachability changes. The server will deem the client's interest in a prefix to have ceased when it does not hear any L-ARP queries for some configured timeout period.

3.3. Client-Server Synchronization

In an L-ARP reply, the server communicates several pieces of information to the client: its hardware address, the MPLS label, Entropy Label capability and metric. Since ARP is a stateless protocol, it is possible that one of these changes without the client knowing, which leads to a loss of synchronization between the client and the server. This loss of synchronization can have several bad effects

If the server's hardware address changes or the MPLS label is repurposed by the server for a different purpose, then packets may be sent to the wrong destination. The consequences can range from suboptimally routed packets to dropped packets to packets being delivered to the wrong customer, which may be a security breach. This last may be the most troublesome consequence of loss of synchronization.

If a destination transitions from entropy label capable to entropy label incapable (an unlikely event) without the client knowing, then packets encapsulated with entropy labels will be dropped. A transition in the other direction is relatively benign.

If the metric changes without the client knowing, packets may be suboptimally routed. This may be the most benign consequence of loss of synchronization.

3.4. Applicability

L-ARP can be used between a host and its Top-of-Rack switch in a Data Center. L-ARP can also be used between a DSLAM and its aggregation switch going to the B-RAS. More generally, L-ARP can be used between an "access node" and its first hop MPLS-enabled device in the context of Seamless MPLS [reference]. In all these cases, L-ARP can handle the presence of multiple connections between the access device and its first hop devices.

ARP is not a routing protocol. The use of L-ARP should be limited to cases where the L-ARP client has a small number of one-hop connections to L-ARP servers. The presence of a complex topology between the L-ARP client and server suggests the use of a different protocol.

3.5. Backward Compatibility

Since L-ARP uses a new hardware type, it is backward compatible with "regular" ARP. ARP servers and clients MUST be able to send out, receive and process ARP messages based on hardware type. They MAY choose to ignore requests and replies of some hardware types; they MAY choose to log errors if they encounter hardware types they do not recognize; however, they MUST handle all hardware types gracefully. For hardware types that they do understand, ARP servers and clients MUST handle operation codes gracefully, processing those they understand, and ignoring (and possibly logging) others.

4. For Future Study

The L-ARP specification is quite simple, and the goal is to keep it that way. However, inevitably, there will be questions and features that will be requested. Some of these are:

1. Keeping L-ARP clients and servers in sync. In particular, dealing with:
 - A. client and/or server restart

- B. lost packets
- C. timeouts
- 2. Withdrawing a response.
- 3. Dealing with scale.
- 4. If there are many servers, which one to pick?
- 5. How can a client make best use of underlying ECMP paths?
- 6. and probably many more.

In all of these, it is important to realize that, whenever possible, a solution that places most of the burden on the server rather than on the client is preferable.

5. L-ARP Message Format

```

 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
+-----+-----+-----+-----+-----+-----+-----+-----+
|               ar$hrd               |               ar$pro               |
+-----+-----+-----+-----+-----+-----+-----+-----+
|   ar$hln   |   ar$pln   |               ar$op               |
+-----+-----+-----+-----+-----+-----+-----+-----+
//                               ar$sha (variable...)                               //
+-----+-----+-----+-----+-----+-----+-----+-----+
//                               ar$spa (variable...)                               //
+-----+-----+-----+-----+-----+-----+-----+-----+
//                               ar$tha (variable...)                               //
+-----+-----+-----+-----+-----+-----+-----+-----+
//                               ar$tpa (variable...)                               //
+-----+-----+-----+-----+-----+-----+-----+-----+

```

Figure 2: L-ARP Packet Format

ar\$hrd Hardware Type: MPLS-over-Ethernet. The value of the field used here is [HTYPE-MPLS-TBD]. To start with, we will use the experimental value HW_EXP2 (256)

ar\$pro Protocol Type: IPv4/IPv6. The value of the field used here is 0x0800 to resolve an IPv4 address and 0x86DD to resolve an IPv6 address.

ar\$hln Hardware Length: the value of the field used here is 12.

ar\$pln Protocol Address Length: for an IPv4 address, the value is 4;
for an IPv6 address, it is 16.

ar\$op Operation Code: set to 1 for request, 2 for reply, and 10 for
ARP-NAK. Other op codes may be used, but this is not anticipated
at this time.

ar\$sha Source Hardware Address: In an L-ARP query message, Source
Hardware Address is irrelevant, and set to all-zeroes. In an
L-ARP reply message, the address follows the 'hardware address'
format specified below.

ar\$spa Source Protocol Address: In an L-ARP query message, this
field carries the sender's IP address. In an L-ARP reply
message, this field carries the target protocol address received
in the corresponding query message.

ar\$tha Target Hardware Address: This field is invalid in both
request and reply messages.

ar\$tpa Target Protocol Address: In an L-ARP query message, this
field carries the IP address for which the client is seeking an
MPLS label. In an L-ARP reply message, this field carries the
Source Protocol Address received in the corresponding L-ARP
query.

Figure 3 describes the format of 'Hardware Address' carried in L-ARP.

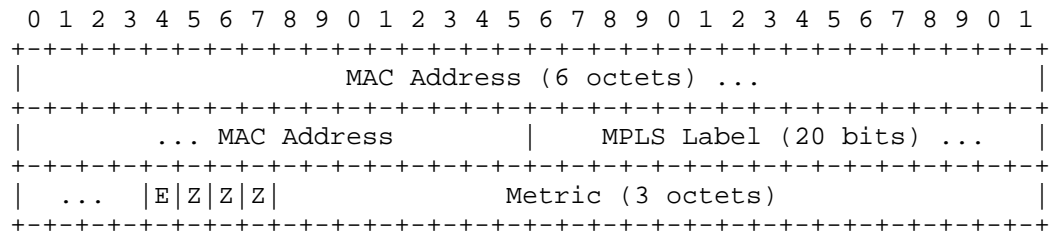


Figure 3: MPLS Hardware Address Format

MAC Address This field contains the Ethernet hardware address that
data packets should be directed to.

MPLS Label This field contains the MPLS label allocated by the
server. This field is valid only in an L-ARP request message.
This field is 20 bits wide, left-justified.

E-bit: Entropy Capability

This field indicates whether the label stack of MPLS data packets sent with the label in this advertisement can contain Entropy Label or not. If this flag is set, the client has the option of inserting ELI and EL as specified in [RFC6790]. The client can choose not to insert ELI/EL pair, if it does not support Entropy Labels, or the local policy does not permit the client to insert ELI/EL. If this flag is clear, the client must not insert ELI/EL into the label stack when sending packets with the advertised L-ARP label.

- Z These bits are not used, and SHOULD be set to zero on sending and ignored on receipt.

If other parameters are deemed useful in the L-ARP reply, they will be added as needed.

6. Security Considerations

TODO

7. IANA Considerations

TODO

8. Acknowledgments

Many thanks to Shane Amante for his detailed comments and suggestions. Many thanks to the team in Juniper prototyping this work for their suggestions on making this variant workable in the context of existing ARP implementations. Thanks too to Luyuan Fang, Alex Semenyaka and Dmitry Afanasiev for their comments and encouragement.

9. Normative References

- [RFC0826] Plummer, D., "Ethernet Address Resolution Protocol: Or converting network protocol addresses to 48.bit Ethernet address for transmission on Ethernet hardware", STD 37, RFC 826, November 1982.
- [RFC2002] Perkins, C., "IP Mobility Support", RFC 2002, October 1996.
- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

[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, Inc.
1194 N. Mathilda Avenue
Sunnyvale, CA 94089
USA

Email: kireeti.kompella@gmail.com

Balaji Rajagopalan
Juniper Networks, Inc.
Prestige Electra, Exora Business Park
Marathahalli - Sarjapur Outer Ring Road
Bangalore 560103
India

Email: balajir@juniper.net

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: December 19, 2014

Z. Li
M. Chen
Huawei
G. Mirsky
Ericsson
June 17, 2014

Routing Extensions for Discovery of Role-based MPLS Label Switching
Router (MPLS LSR) Traffic Engineering (TE) Mesh Membership
draft-li-ccamp-role-based-automesh-02

Abstract

A Traffic Engineering (TE) mesh-group is defined as a group of Label Switch Routers (LSRs) that are connected by a full mesh of TE LSPs. Routing (OSPF and IS-IS) extensions for discovery Multiprotocol Label Switching (MPLS) LSR TE mesh membership has been defined to automate the creation of mesh of TE LSPs.

This document introduces a role-based TE mesh-group that applies to the scenarios where full mesh TE LSPs is not necessary and TE LSPs setup depends on the roles of LSRs in a TE mesh-group. Interior Gateway Protocol (IGP) routing extensions for automatic discovery of role-based TE mesh membership are defined accordingly.

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 December 19, 2014.

Copyright Notice

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

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

Table of Contents

1. Introduction	2
1.1. Terminology	3
2. Role-based TE Mesh Group	4
3. IGP Role-based TE Mesh-group Extensions	4
3.1. OSPF TE-ROLE-MESH-GROUP TLV Format	4
3.2. IS-IS TE-ROLE-MESH-GROUP Sub-TLV Format	7
4. Elements of Procedure	10
4.1. OSPF	10
4.2. IS-IS	11
5. Backward Compatibility	12
6. IANA Considerations	12
6.1. OSPF	12
6.2. IS-IS	13
7. Security Considerations	13
8. Acknowledgements	13
9. References	13
9.1. Normative References	13
9.2. Informative References	14
Authors' Addresses	14

1. Introduction

A TE mesh-group [RFC4972] is defined as a group of LSRs that are connected by a full mesh of TE LSPs. [RFC4972] specifies Intermediate System-to-Intermediate System (IS-IS) and Open Shortest Path First (OSPF) extensions to provide an automatic discovery of the set of LSR members of a TE mesh-group in order to automate the creation of such mesh of TE LSPs. This is called "auto-mesh TE" or "auto-mesh". The auto-mesh TE significantly simplifies the configuration and deployment of TE LSPs.

In some scenarios, it may not be necessary to establish full mesh TE LSPs among all the LSRs of a TE mesh-group. An example of the use case of non-full mesh of TE LSPs in the mobile backhaul (MBH) networks is presented in ([I-D.li-mpls-seamless-mpls-mbb]). In MBH network TE LSPs are usually setup between the Cell Site Gateways(CSGs) and the Radio Network Controller (RNC) Site Gateways(RSGs). TE LSPs interconnecting CSGs and TE LSPs interconnecting RSGs are not necessary. In most deployments the number of CSGs is very large and there are much more CSGs than RSGs in an MBH domain. With the auto-mesh mechanism defined[RFC4972] full mesh of TE LSPs will be established interconnecting CSGs and RSGs. As result large number of unnecessary TE LSPs will be established interconnecting CSGs and interconnecting RSGs. This likely will not scale well with addition of more CSG devices, would stress control plane with unwarranted RSVP state.

Thus there are requirements to optimize the auto-mesh TE and to reduce the number of unnecessary TE LSPs. This document introduces a "role-based auto-mesh TE" or "role-based auto-mesh" where the setup of the TE LSPs is based on the role of the LSRs within a particular TE mesh-group. Therefore, besides the discovery of the membership of a TE mesh-group, it needs to discover the role of each node in the TE mesh-group.

Another scenario to which the role-based auto-mesh TE can apply is the Resource Reservation Protocol-Traffic Engineering (RSVP-TE) Point-to-Multipoint (P2MP) TE LSP[RFC4875] scenario. For a RSVP-TE P2MP TE LSP, the root LSR has to know all the leaf LSRs before signalling the P2MP TE LSP. The automatic discovery mechanisms defined in this document can be used to discover the leaf LSRs for P2MP TE LSPs.

This document defines IGP routing extensions to automatically discover of the members and their roles of a TE mesh-group.

1.1. Terminology

RSVP-TE - Resource Reservation Protocol-Traffic Engineering

P2MP - Point-to-Multipoint

IS-IS - Intermediate System-to-Intermediate System

OSPF - Open Shortest Path First

CSG - Cell Site Gateway

RNC - Radio Network Controller

MBH - Mobile Backhaul

MPLS - Multiprotocol Label Switching

LSP - Label Switched Path

TE LSP - Traffic Engineered LSP

2. Role-based TE Mesh Group

A role-based TE mesh-group is a special TE mesh-group where TE LSPs will not be established among all member LSRs. In a role-based TE mesh-group LSRs will have different roles. TE LSPs setup depends on the roles of the LSRs in a TE mesh-group. This document introduces two types of role-based TE mesh group: Hub-Spoke and Root-Leaf.

For a Hub-Spoke TE mesh-group, an LSR can be a Hub, Spoke or both Hub and Spoke LSR in a group. The rules for Hub-Spoke TE mesh-group are as follows:

TE LSPs SHOULD only be setup between Spoke and Hub LSRs.

TE LSPs MUST NOT be setup between/among Spoke LSRs.

TE LSPs MUST NOT be setup between/among Hub LSRs.

For a Root-Leaf TE mesh-group, an LSR can be a Root, a Leaf or both a Root and Leaf LSR. Once the membership and roles are determined, the root LSRs can signal the P2MP TE LSPs toward all the Leaf LSRs. There may be multiple P2MP TE LSPs within a TE mesh-group.

3. IGP Role-based TE Mesh-group Extensions

3.1. OSPF TE-ROLE-MESH-GROUP TLV Format

The OSPF TE-ROLE-MESH-GROUP TLV is used to advertise that an LSR joins/leaves a role-based TE mesh-group and the role of the LSR in the TE mesh-group. The OSPF TE-ROLE-MESH-GROUP TLV format for IPv4 (Figure 2) and IPv6 (Figure 3) is as follows:

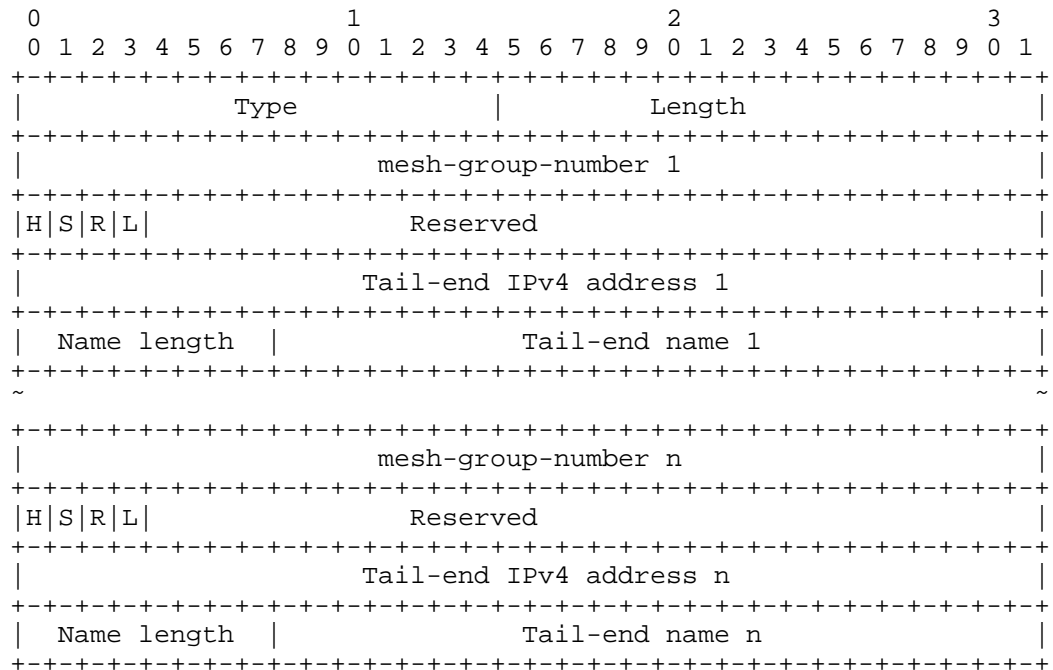


Figure 2 - OSPF TE-ROLE-MESH-GROUP TLV format (IPv4 Address)

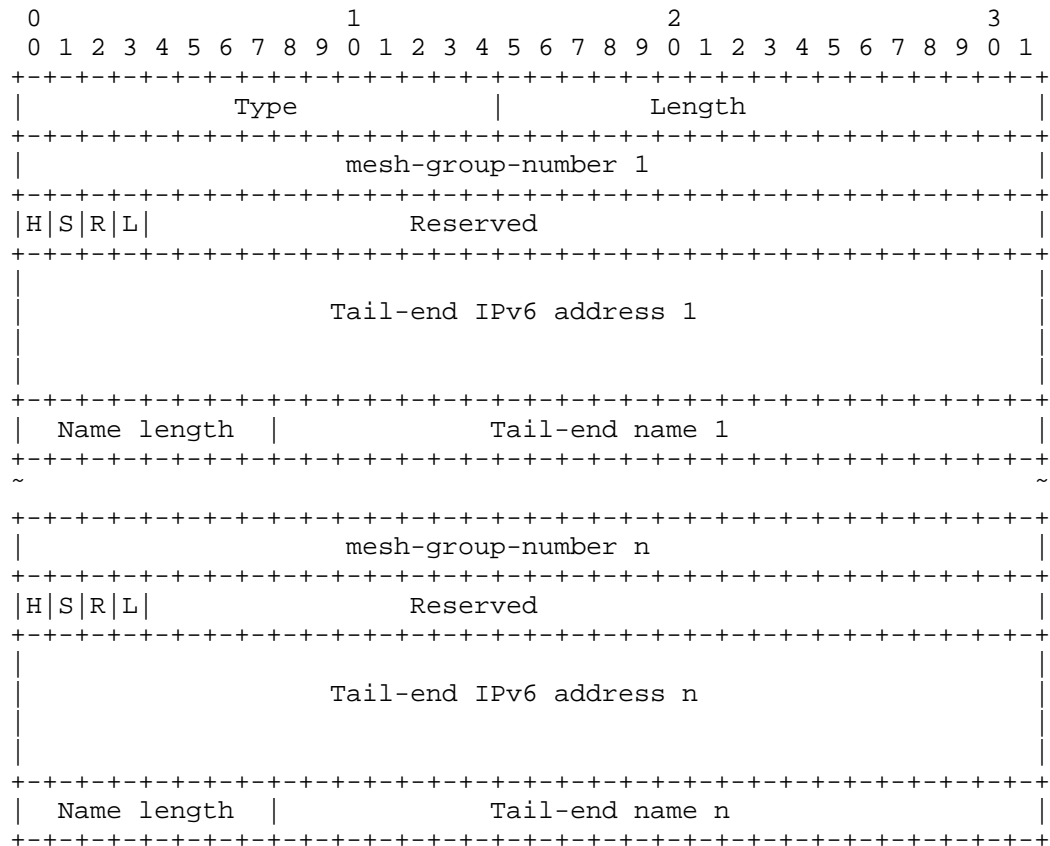


Figure 3 - OSPF TE-ROLE-MESH-GROUP TLV format (IPv6 Address)

The Type of OSPF TE-ROLE-MESH-GROUP TLV for IPv4 is TBD1, the value of the Length is variable.

The Type of OSPF TE-ROLE-MESH-GROUP TLV for IPv6 is TBD2, the value of the Length is variable.

The OSPF TE-ROLE-MESH-GROUP TLV may contain one or more role-based mesh-group entries. Each entry corresponds to a role-based TE mesh-group. The definition of the mesh-group-number, the Tail-end address, the Name length and the Tail-end name in each role-based mesh group entry is the same as that of OSPF TE-MESH-GROUP TLV defined in [RFC4972].

In addition, for each mesh group entry, an four-octet flag field is introduced and four flags are defined in this document. Other bits

are reserved for future use and MUST be set to zero when sent, and MUST be ignored when received.

The H (Hub) bit, when set, it indicates the LSR is a Hub LSR.

The S (Spoke) bit, when set, it indicates the LSR is a Spoke LSR.

The R (Root) bit, when set, it indicates an LSR is a Root LSR.

The L (Leaf) bit, when set, it indicates an LSR is a Leaf.

The H and S bit are dedicated for Hub-Spoke TE mesh-group and can be both set. When both bits set, it indicates that an LSR has both the Hub and Spoke role in the group.

The R and Leaf bit can be both set, when both bits set, it indicates an LSR is a Root and Leaf LSR. The R bit and Leaf bit are only used for Root-Leaf TE mesh-group, for other TE mesh-groups, it MUST be set to zero and MUST be ignored when received.

3.2. IS-IS TE-ROLE-MESH-GROUP Sub-TLV Format

The IS-IS TE-ROLE-MESH-GROUP sub-TLV is used to advertise that an LSR joins/leaves a TE mesh-group and the role of the LSR in the TE mesh-group. The IS-IS TE-ROLE-MESH-GROUP sub-TLV format for IPv4 (Figure 4) and IPv6 (Figure 5) is as follows:

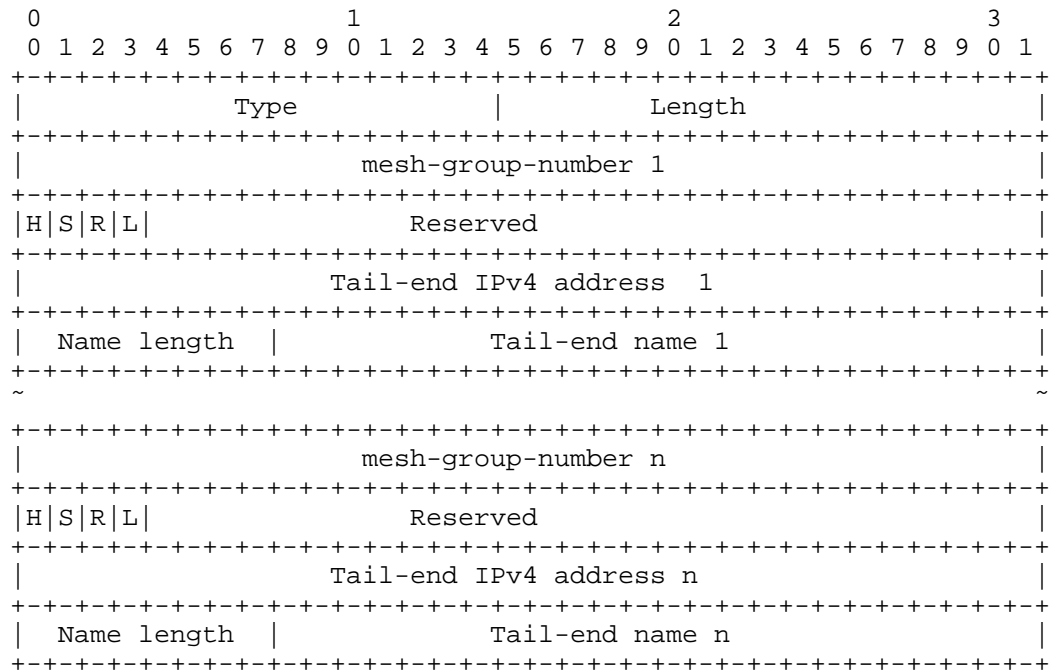


Figure 4 - IS-IS TE-ROLE-MESH-GROUP sub-TLV format (IPv4 Address)

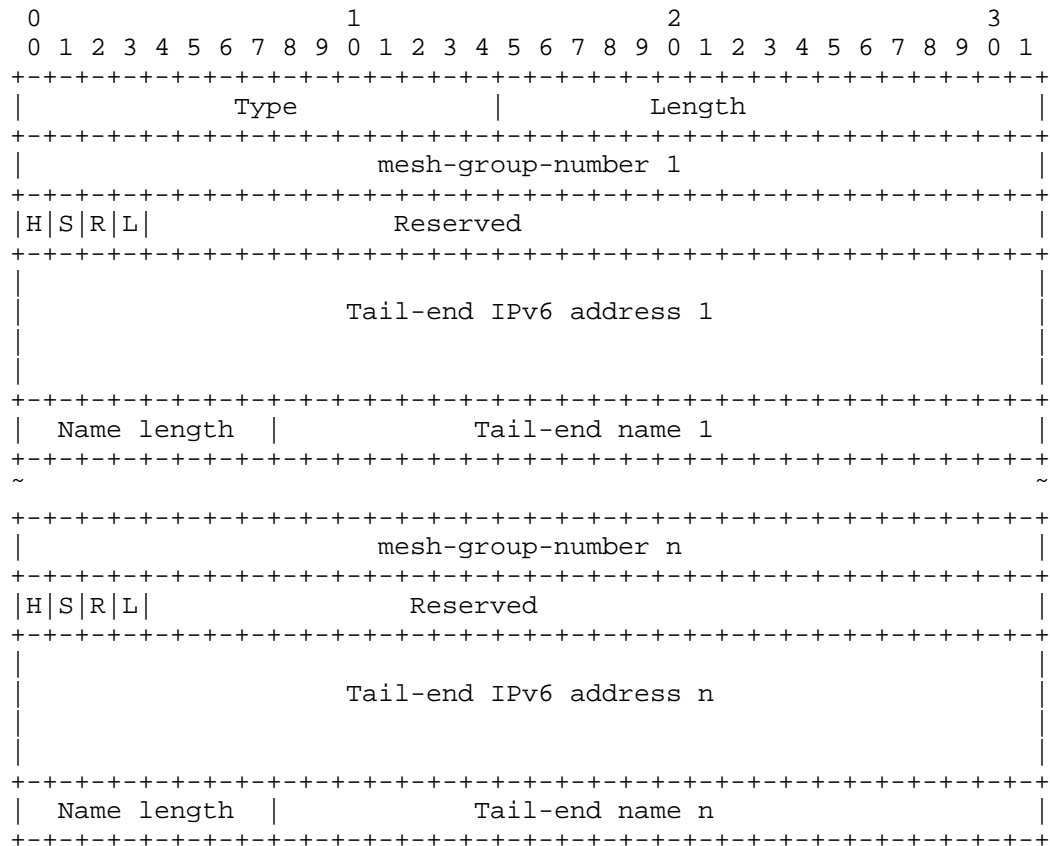


Figure 5 - IS-IS TE-ROLE-MESH-GROUP sub-TLV format (IPv6 Address)

The Type of IS-IS TE-ROLE-MESH-GROUP sub-TLV for IPv4 is TBD3, the value of the Length is variable.

The Type of IS-IS TE-ROLE-MESH-GROUP sub-TLV for IPv6 is TBD4, the value of the Length is variable.

The IS-IS Role-based TE-ROLE-MESH-GROUP sub-TLV may contain one or more role-based mesh-group entries. Each entry corresponds to a role-based TE mesh-group. The definition of the fields, mesh-group-number, Tail-end address, Name length and Tail-end name in each role-based mesh group entry is the same as that of IS-IS TE-MESH-GROUP sub-TLV defined in [RFC4972].

The H, S, R and L bits are defined as in Section 3.1 of this document.

4. Elements of Procedure

The OSPF TE-ROLE-MESH-GROUP TLV is carried within the OSPF Routing Information LSA, and the IS-IS TE-ROLE-MESH-GROUP sub-TLV is carried within the IS-IS Router capability TLV. As such, elements of procedure are inherited from those defined in [RFC4970] and [RFC4971] for OSPF and IS-IS respectively. Specifically, a router **MUST** originate a new LSA/LSP whenever the content of this information changes, or whenever required by regular routing procedure (e.g., updates).

The TE-ROLE-MESH-GROUP TLV is **OPTIONAL** and **MUST NOT** include more than one of each of the IPv4 instances or the IPv6 instance. If either the IPv4 or the IPv6 OSPF TE-ROLE-MESH-GROUP TLV occurs more than once within the OSPF Router Information LSA, only the first instance is processed, subsequent TLV(s) **MUST** be ignored. Similarly, if either the IPv4 or the IPv6 IS-IS TE-ROLE-MESH-GROUP sub-TLV occurs more than once within the IS-IS Router capability TLV, only the first instance is processed, subsequent TLV(s) **MUST** be ignored.

4.1. OSPF

The TE-ROLE-MESH-GROUP TLV is advertised within an OSPF Router Information opaque LSA (opaque type of 4, opaque ID of 0) for OSPFv2 [RFC2328] and within a new LSA (Router Information LSA) for OSPFv3 [RFC5340]. The Router Information LSAs for OSPFv2 and OSPFv3 are defined in [RFC4970].

A router **MUST** originate a new OSPF router information LSA whenever the content of any of the advertised TLV changes or whenever required by the regular OSPF procedure (LSA update (every LSRefreshTime)). If an LSR desires to join or leave a particular role-based TE mesh group or an LSR desires to change its role in a mesh group, it **MUST** originate a new OSPF Router Information LSA comprising the updated TE-ROLE-MESH-GROUP TLV. In the case of a join, a new entry will be added to the TE-ROLE-MESH-GROUP TLV; if the LSR leaves a mesh-group, the corresponding entry will be removed from the TE-ROLE-MESH-GROUP TLV; if the LSR changes its role in the role-based mesh group, the corresponding entry will be updated in the TE-ROLE-MESH-GROUP TLV. Note that these operations can be performed in the context of a single LSA update. An implementation **SHOULD** be able to detect any change to a previously received TE-ROLE-MESH-GROUP TLV from a specific LSR.

As defined in [RFC5250] for OSPFv2 and in [RFC5340] for OSPFv3, the flooding scope of the Router Information LSA is determined by the LSA Opaque type for OSPFv2 and the values of the S1/S2 bits for OSPFv3.

For OSPFv2 Router Information opaque LSA:

- Link-local scope: type 9;
- Area-local scope: type 10;
- Routing-domain scope: type 11. In this case, the flooding scope is equivalent to the Type 5 LSA flooding scope.

For OSPFv3 Router Information LSA:

- Link-local scope: OSPFv3 Router Information LSA with the S1 and S2 bits cleared;
- Area-local scope: OSPFv3 Router Information LSA with the S1 bit set and the S2 bit cleared;
- Routing-domain scope: OSPFv3 Router Information LSA with S1 bit cleared and the S2 bit set.

A router may generate multiple OSPF Router Information LSAs with different flooding scopes.

The Role-based TE-MESH-GROUP TLV may be advertised within an Area-local or Routing-domain scope Router Information LSA, depending on the MPLS TE mesh group profile:

- If the MPLS TE mesh-group is contained within a single area (all the LSRs of the mesh-group are contained within a single area), the TE-ROLE-MESH-GROUP TLV MUST be generated within an Area-local Router Information LSA.
- If the MPLS TE mesh-group spans multiple OSPF areas, the TE Role-based mesh-group TLV MUST be generated within a Routing-domain scope router information LSA.

When the router receives TE-ROLE-MESH-GROUP TLV, it SHOULD setup MPLS TE LSPs according rules which defined in the Section 3.

4.2. IS-IS

The TE-ROLE-MESH-GROUP sub-TLV is advertised within the IS-IS Router CAPABILITY TLV defined in [RFC4971].

An IS-IS router MUST originate a new IS-IS LSP whenever the content of any of the advertised sub-TLV changes or whenever required by regular IS-IS procedure (LSP updates). If an LSR desires to join or leave a particular role-based TE mesh group or an LSR desires to

change its role in a mesh group, it MUST originate a new LSP comprising the refreshed IS-IS Router capability TLV comprising the updated TE-ROLE-MESH-GROUP sub-TLV. In the case of a join, a new entry will be added to the TE-ROLE-MESH-GROUP sub-TLV; if the LSR leaves a mesh-group, the corresponding entry will be deleted from the TE-ROLE-MESH-GROUP sub-TLV; if the LSR changes its role in the role-based mesh group, the corresponding entry will be updated in the TE-ROLE-MESH-GROUP sub-TLV. Note that these operations can be performed in the context of a single update. An implementation SHOULD be able to detect any change to a previously received TE-ROLE-MESH-GROUP sub-TLV from a specific LSR.

If the flooding scope of a TE-ROLE-MESH-GROUP sub-TLV is limited to an IS-IS level/area, the sub-TLV MUST NOT be leaked across level/area and the S flag of the Router CAPABILITY TLV MUST be cleared. Conversely, if the flooding scope of a TE-ROLE-MESH-GROUP sub-TLV is the entire routing domain, the TLV MUST be leaked across IS-IS levels/areas, and the S flag of the Router CAPABILITY TLV MUST be set. In both cases, the flooding rules specified in [RFC4971] apply.

As specified in [RFC4971], a router may generate multiple IS-IS Router CAPABILITY TLVs within an IS-IS LSP with different flooding scopes.

When the router receives TE-ROLE-MESH-GROUP sub-TLV, it SHOULD setup MPLS TE LSPs according rules which defined in the Section 3.

5. Backward Compatibility

For a role-based TE mesh-group, if there are some LSRs only supporting mechanisms defined [RFC4972], all the LSRs of the mesh-group MUST process as defined in [RFC4972]. The operators should avoid to add an LSR that does not support role-based auto-mesh TE to a role-based TE mesh-group.

6. IANA Considerations

6.1. OSPF

The registry for the Router Information LSA is defined in [RFC4970]. IANA assigned a new OSPF TLV code-point for the TE-ROLE-MESH-GROUP TLVs carried within the Router Information LSA.

Value	TLV	References
-----	-----	-----
TBD1	TE-ROLE-MESH-GROUP TLV (IPv4)	this document
TBD2	TE-ROLE-MESH-GROUP TLV (IPv6)	this document

6.2. IS-IS

The registry for the Router Capability TLV is defined in [RFC4971]. IANA assigned a new IS-IS sub-TLV code-point for the TE-ROLE-MESH-GROUP sub-TLVs carried within the IS-IS Router Capability TLV.

Value -----	Sub-TLV -----	References -----
TBD3	TE-ROLE-MESH-GROUP sub-TLV (IPv4)	this document
TBD4	TE-ROLE-MESH-GROUP sub-TLV (IPv6)	this document

7. Security Considerations

The function described in this document does not create any new security issues for the OSPF and IS-IS protocols, the security considerations described in [RFC4972] apply here.

8. Acknowledgements

The authors would like to thank Loa Andersson for his valuable comments.

The authors would also like to thank the authors of [RFC4972] from where we have taken most of the elements procedures.

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.
- [RFC2328] Moy, J., "OSPF Version 2", STD 54, RFC 2328, April 1998.
- [RFC4970] Lindem, A., Shen, N., Vasseur, JP., Aggarwal, R., and S. Shaffer, "Extensions to OSPF for Advertising Optional Router Capabilities", RFC 4970, July 2007.
- [RFC4971] Vasseur, JP., Shen, N., and R. Aggarwal, "Intermediate System to Intermediate System (IS-IS) Extensions for Advertising Router Information", RFC 4971, July 2007.
- [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.

- [RFC5250] Berger, L., Bryskin, I., Zinin, A., and R. Coltun, "The OSPF Opaque LSA Option", RFC 5250, July 2008.
- [RFC5340] Coltun, R., Ferguson, D., Moy, J., and A. Lindem, "OSPF for IPv6", RFC 5340, July 2008.

9.2. Informative References

- [I-D.li-mpls-seamless-mpls-mbb]
Li, Z., Li, L., Morillo, M., and T. Yang, "Seamless MPLS for Mobile Backhaul", draft-li-mpls-seamless-mpls-mbb-01 (work in progress), February 2014.
- [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.

Authors' Addresses

Zhenbin Li
Huawei

Email: lizhenbin@huawei.com

Mach(Guoyi) Chen
Huawei

Email: mach.chen@huawei.com

Greg Mirsky
Ericsson

Email: gregory.mirsky@ericsson.com

Network Working Group
Internet-Draft
Intended status: Informational
Expires: January 4, 2015

Z. Li
Q. Zhao
X. Chen
Huawei Technologies
T. Yang
China Mobile
R. Raszuk
Individual
July 3, 2014

A Framework of MPLS Global Label
draft-li-mpls-global-label-framework-02

Abstract

The document defines the framework of MPLS global label including the label allocation method for MPLS global label, the representation of MPLS global label and the process of control plane and data plane for MPLS global label.

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 4, 2015.

Copyright Notice

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

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

Table of Contents

1. Introduction	3
2. Terminology	3
3. MPLS Global Label Allocation Methods	3
3.1. Special-Purpose MPLS Label	3
3.2. Domain Wide Labels	3
3.2.1. Label Allocation Methods	4
4. Representation of MPLS Global Label	4
4.1. Per-platform Label Space	4
4.2. Context-Specific Label Space	5
5. Control Plane for MPLS Global Label	5
5.1. Architecture	5
5.2. In-Band Global Label Allocation	7
5.2.1. Label Allocation in Per-Platform MPLS Label Space . .	7
5.2.2. Label Allocation in Context-Specific Label Space . .	9
5.3. Label Mapping Distribution	9
5.4. Inter-Domain Label Negotiation	9
5.5. Protocol Extensions Requirement	10
5.5.1. IGP Protocol Extensions	10
5.5.2. BGP Protocol Extensions	10
5.5.3. PECP Protocol Extensions	10
6. Data Plane of MPLS Global Label	11
6.1. Global Label in Per-Platform Label Space	11
6.2. Global Label in Context-Specific Label Space	11
6.3. Global Process of Inner Global Label	11
7. IANA Considerations	12
8. Security Considerations	12
9. References	12
9.1. Normative References	12
9.2. Informative References	13
Authors' Addresses	13

1. Introduction

[I-D.li-mpls-global-label-usecases] proposes possible usecases of MPLS global label. MPLS global label can be used for identification of the location, the service and the network in different application scenarios.

Several MPLS global label allocation mechanisms has been proposed in [RFC5331], [I-D.raszuk-mpls-domain-wide-labels], etc.. This document is to define the framework for MPLS global label based on the existing work and more emerging applications. The framework includes the label allocation method for MPLS global label, the representation of MPLS global label and the process of control plane and data plane for MPLS global label.

2. Terminology

FEC: Forward Equivalence Class

MVPN: Multicast VPN

PCE: Path Computation Element

SRGB: Global Segment Routing Block

3. MPLS Global Label Allocation Methods

MPLS global label is the label which meaning can be understood by all nodes or part of nodes in the network. These nodes can be nodes in one domain or nodes spanning multiple domains.

3.1. Special-Purpose MPLS Label

Special-purpose MPLS label defined in [RFC7274] is a type of special global label. These labels have specific well-known meaning which can be understood and processed accordingly by all MPLS nodes in the network. These labels are allocated and retired by IANA. How to allocate and retire these labels is specified in [RFC7274].

3.2. Domain Wide Labels

Besides the special-purpose labels which have the global meaning and are defined by the IANA, it is necessary to provide dynamic allocation mechanisms to allocate global labels to satisfy requirements of emerging possible applications ([I-D.li-mpls-global-label-usecases]). Such global labels may be not possible to be understood by all network nodes like the special-purpose label. That is, these labels may be only understood by all

nodes or part of nodes in one domain or multiple domains. This type of global label can also be called as Domain Wide Label. The scope of domains for Domain Wide Label is service-specific or management-specific which is out of scope of this document.

Note: In the following sections of this document, the global label always means Domain Wide Label. That is, the global label and the Domain Wide Label have the same meaning.

3.2.1. Label Allocation Methods

There are two types of label allocation methods for Domain Wide Labels: out-of-band label allocation and in-band label allocation.

Out-of-band label allocation means that the global labels are planned and designated manually for special usage. The typical scenario is Segment Routing. When MPLS is applied for Segment Routing, the global labels allocated for node segments is based on the reserved SRGB and the designated unique Segment ID. In essence the global uniqueness of these label is guaranteed by manual planning. So this method can be seen as the out-of-band label allocation.

In-band label allocation means that the global labels are requested and allocated dynamically through control protocols in the domains. The typical example is the upstream MPLS label assignment defined in [RFC5331]. The method has been adopted in BGP-based MVPN ([RFC6514]) in which the root PE allocates labels to represent MVPN instances and advertise the label binding to leaf PEs for the scenario that multiple MVPNs shares one P-multicast tree.

Choice of the two methods is related with scalability of the possible applications. If the scale of the application is limited, the out-of-band method is enough. Otherwise, the in-band method must be taken into account.

4. Representation of MPLS Global Label

4.1. Per-platform Label Space

The labels in the per-platform label space can be used for Domain Wide label. The advantage of this method is that the existing MPLS forward plane which is used for widely deployed applications based on the per-platform label can be reused well to support global labels. The challenge for the method is that the existing MPLS protocols such as LDP, BGP and RSVP-TE are always allocate local labels from the space which may cause the confliction with the global label allocation. This confliction could be prevented through division of

the per-platform space into multiple segments used for local label and global label respectively.

4.2. Context-Specific Label Space

The concept of Context-Specific Label Space is defined in [RFC5331]. The labels in Context-Specific label space can also be used for Domain Wide label. The Context-Specific label space is isolated from the per-platform label space and the confliction issue of label allocation can be avoided naturally. The challenge for the method is the possible complexity of both control plane and forward plane introduced by multiple label spaces management.

If the Context-Specific label space is used for global labels, it is necessary to determine the Context Identifier for the label space. There are two methods as follows:

-- Service-specific Context Identifier: The Context Identifier is determined by the service. For example, in [RFC6514], the Tunnel-Specific label space is introduced in which the P-Tunnel Identifier becomes the Context Identifier for the label space.

-- MPLS Global Label Indicator: It is to define a well-known Context-Specific label space for global label. The label space is indicated by the MPLS Global Label Indicator which can be seen as a well-known Context Identifier. In the forwarding plane, the MPLS Global Label Indicator is a special-purpose label to indicate that next label in the MPLS label stack of each transported packet is Domain Wide Labels. The value of the special-purpose label needs to be allocated by IANA according to [RFC7274].

5. Control Plane for MPLS Global Label

5.1. Architecture

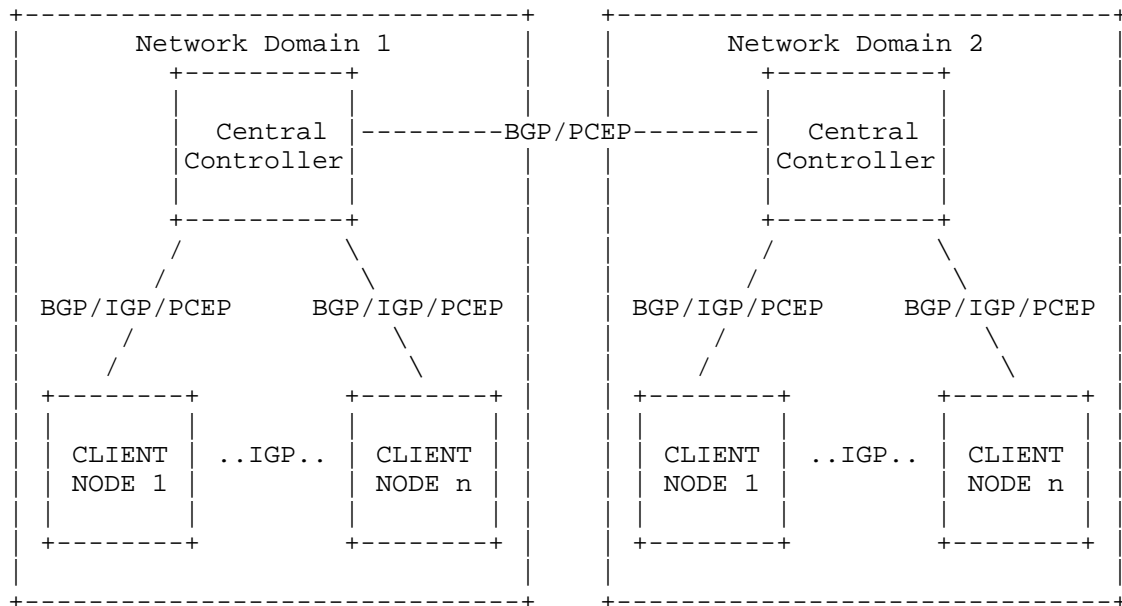


Figure 1: Architecture of In-band Domain-Wide Label Allocation

MPLS global label should be allocated centrally to guarantee all nodes can understand the same meaning for a specific global label. It is natural to adopt the central control architecture for the in-band label allocation. In the architecture the central controller is responsible for allocating the global labels and advertising to the client nodes in the network. When client nodes receives the label binding, it will install the corresponding forwarding entry for the global label.

The applications based on global labels are different: they may need advertise global label to all nodes of a domain, edge nodes of a domain or part of nodes of a domain. IGP extensions, BGP extensions and PCEP extensions are appropriate for these applications respectively. In addition, the global label may be negotiated across multiple domains, it will adopt BGP extensions and PCEP extensions.

Central Control of global labels is the logical functionality which can be deployed in the independent server or in the network device. For example, the upstream label assignment for BGP-based MVPN is done by the root node of MVPN which can be seen as the central controller for the global label.

5.2. In-Band Global Label Allocation

5.2.1. Label Allocation in Per-Platform MPLS Label Space

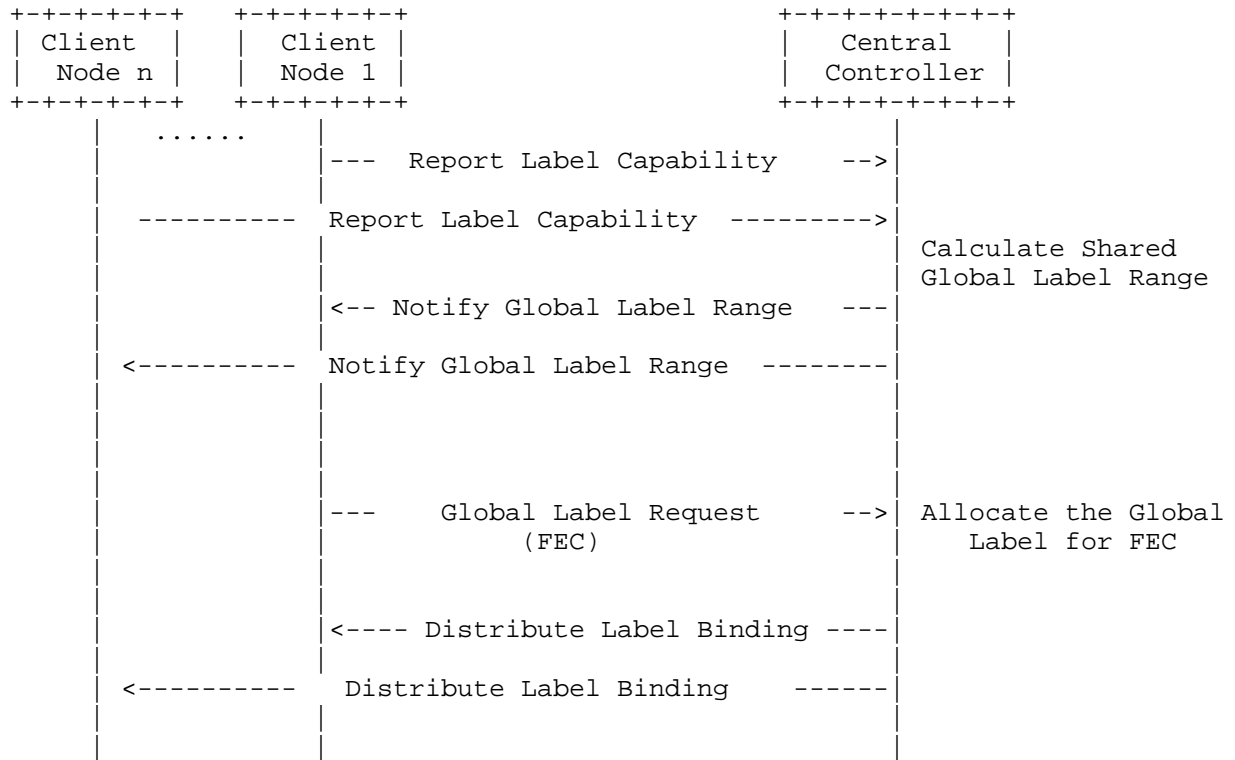


Figure 2: Procedures of Global Label Allocation

Procedures of global label allocation from per-platform label space is shown in the Figure 2. There are two import phases for these procedures: Shared MPLS global label range calculation and MPLS global label allocation.

5.2.1.1. Shared MPLS Global Label Range Calculation

1. Clients nodes should report MPLS label capability to the central controller.
2. The central controller collects MPLS label capability of all nodes. Then it can calculate the shared MPLS global label range for all nodes.

3. The central controller should notify the shared global label range to all client nodes.

Report of label capability and notification of shared MPLS global range can be done by IGP, BGP or PCEP extensions.

5.2.1.2. Label Allocation

There are two methods for the global label allocation: On-demand label allocation and Unsolicited label allocation.

1. On-demand allocation

This method is that the global label allocation is done by the central controller based on the label requirement from client nodes. The procedures of on-demand allocation are as follows:

- 1) The client node should send the global label request for specific usage to the central controller. FEC(Forward Equivalence Class) should be incorporated in the MPLS global label request message.
- 2) When the central controllers receives the MPLS global label request, it should allocate the label from the shared MPLS global label range of all nodes.
- 3) The central controller distributes the MPLS global label mapping message to all client nodes. Thus the MPLS global label for specific usage can be understood by all client nodes.
- 4) The client nodes receive the MPLS global label mapping message and install the corresponding MPLS forwarding entry for the global label.

Label request and distribution of label mapping which are used in on-demand allocation can be done by BGP extensions or PCEP extensions.

2. Unsolicited allocation

This method is that the central controller directly allocates the global label without receiving the label request. The procedures of unsolicited allocation are as follows:

- 1) Discovery of service: this can be implemented by configuration or auto discovery which is service-specific and out of scope of this document.
- 2) The central controller allocates the global label from the global label space for the service.

3) The central controller distributes the MPLS global label mapping message to all client nodes. Thus the MPLS global label for specific usage can be understood by all related client nodes.

4) The client nodes receive the MPLS global label mapping message and install the corresponding MPLS forwarding entry for the global label.

Distribution of label mapping which is used in unsolicited allocation can be done by IGP extensions, BGP extensions or PCEP extensions.

5.2.2. Label Allocation in Context-Specific Label Space

As mentioned in previous section, there can be two types of Context-Specific label space for global label allocation. For the Context-Specific label space identified by service-specific context identifier, the label allocation procedures are service-specific and these procedures are out of scope of this document. For the Context-Specific label space identified by MPLS Global Label Indicator, since the label space is well-known, it is necessary to calculate the share global label range like the global allocation in the per-platform label space. Except this, other procedures for global label allocation are similar as the global label allocation in per-platform label space.

5.3. Label Mapping Distribution

After allocating the global label by the central controller, the label mapping must be distributed to all involved nodes of the specific global-label-based service. If the central controller connects to all involved nodes, the label mapping can be directly advertised to these nodes. But if the central controller only connects part of the involved nodes, it not only needs to distribute the label mapping to the connected client nodes, but also the label mapping should be distributed to other client nodes by the clients nodes which receive the label mapping from the central controller. The distribution of label mapping among client nodes can be implemented by IGP extensions.

5.4. Inter-Domain Label Negotiation

If the global label for the service needs to be allocated across multiple domains, PCEP extensions or BGP extensions can be introduced for label negotiation across multiple domains.

5.5. Protocol Extensions Requirement

5.5.1. IGP Protocol Extensions

REQ 01. Report Label Capability from client nodes to the central controller.

REQ 02. Notify the shared global label range from the central controller to client nodes.

REQ 03: Distribute label mapping from the central controller to client node.

REQ 04: Distribute label mapping among client nodes.

5.5.2. BGP Protocol Extensions

REQ 11. Report Label Capability from client nodes to the central controller.

REQ 12. Notify the shared global label range from the central controller to client nodes.

REQ 13: Send global label request from client nodes to the central controller.

REQ 14: Distribute label mapping from the central controller to client node.

REQ 15: Inter-domain global label negotiation

5.5.3. PECP Protocol Extensions

REQ 21. Report Label Capability from client nodes to the central controller.

REQ 22. Notify the shared global label range from the central controller to client nodes.

REQ 23: Send global label request from client nodes to the central controller.

REQ 24: Distribute label mapping from the central controller to client node.

REQ 25: Inter-domain global label negotiation

6. Data Plane of MPLS Global Label

6.1. Global Label in Per-Platform Label Space

For global label allocated from the per-platform label space, the existing MPLS forwarding mechanism can be reused without modification.

6.2. Global Label in Context-Specific Label Space

For a global label allocated within the Context-Specific label space, it is necessary to maintain multiple MPLS label forwarding table in the forwarding plane. When forwarding packets with global label encapsulation, it must decapsulate the label for the Context Identifier firstly to determine the MPLS label forwarding table of the corresponding Context-Specific label space. Then it will decapsulate the next label and search the corresponding MPLS forwarding entry in the Context-Specific label space. The encapsulation of the global label from the Context-Specific label space is shown as follows:

```
+-----+
| Global Label | Global Label |
| Indicator   |               |
+-----+
```

6.3. Global Process of Inner Global Label

Because the label forwarding entry for the global label can be created in multiple nodes in the network, there may be one application scenario for which the global label is located in the middle of the label stack of the transported packet and should be processed by all possible node. For example, the Entropy Label for ECMP can be encapsulated multiple times following multiple node segments in Segment Routing. This method may cause the depth of the label stack of the packet is too deep to process. In order to solve this issue, the global label can be introduced to represent the same process of all possible nodes. Thus the depth of the label stack can be reduced. This method can be implemented by introducing a special-purpose label which is named as Global Process Indicator (GPI). When the Global Process Indicator is encapsulated in the packet, it indicates that the next global label SHOULD be process by each node along the path.

The encapsulation of the global label allocated from the per-platform label space which needs to be globally processed is as follows:

Global Process Indicator	Global Label
-----------------------------	--------------

The encapsulation of the global label allocated from the Context-Specific label space indicated by MPLS Global Label Indicator which needs to be globally processed is as follows:

Global Process Indicator	Global Label Indicator	Global Label
-----------------------------	---------------------------	--------------

7. IANA Considerations

Following two special-purpose labels defined in this document needs to be allocated by IANA:

- Global Label Indicator
- Global Process Indicator

8. Security Considerations

TBD.

9. References

9.1. Normative References

- [I-D.li-mpls-global-label-usecases]
Li, Z., Zhao, Q., and T. Yang, "Usecases of MPLS Global Label", draft-li-mpls-global-label-usecases-01 (work in progress), February 2014.
- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC5331] Aggarwal, R., Rekhter, Y., and E. Rosen, "MPLS Upstream Label Assignment and Context-Specific Label Space", RFC 5331, August 2008.
- [RFC7274] Kompella, K., Andersson, L., and A. Farrel, "Allocating and Retiring Special-Purpose MPLS Labels", RFC 7274, June 2014.

9.2. Informative References

- [I-D.raszuk-mpls-domain-wide-labels]
Raszuk, R., "MPLS Domain Wide Labels", draft-raszuk-mpls-domain-wide-labels-01 (work in progress), January 2014.
- [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

Zhenbin Li
Huawei Technologies
Huawei Bld., No.156 Beiqing Rd.
Beijing 100095
China

Email: lizhenbin@huawei.com

Quintin Zhao
Huawei Technologies
125 Nagog Technology Park
Acton, MA 01719
US

Email: quintin.zhao@huawei.com

Xia Chen
Huawei Technologies
Huawei Bld., No.156 Beiqing Rd.
Beijing 100095
China

Email: jescia.chenxia@huawei.com

Tianle Yang
China Mobile
32, Xuanwumenxi Ave.
Beijing 01719
China

Email: yangtianle@chinamobile.com

Robert Raszuk
Individual

Email: robert@raszuk.net

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: January 5, 2015

Z. Li
X. Zeng
Huawei Technologies
July 4, 2014

Proxy MPLS Traffic Engineering Label Switched Path(LSP)
draft-li-mpls-proxy-te-lsp-01

Abstract

This document describes a method to setup MPLS TE proxy egress LSP which can help 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 5, 2015.

Copyright Notice

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

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents

carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction	2
2. Terminology	3
3. Problem Statement	3
4. Solutions	3
5. Proxy Destination Object	5
5.1. Format	5
5.2. Procedures	6
6. IANA Considerations	6
7. Security Considerations	6
8. Acknowledgements	6
9. References	6
9.1. Normative References	7
9.2. Informative References	7
Authors' Addresses	7

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, etc. to build intra-area segments and uses MP-BGP as the inter-area routing and label distribution protocol.

In the typical Seamless MPLS architecture, LDP DoD are adopted to setup the segment LSP which is stitched with BGP LSP. When Seamless MPLS is applied to integrate the mobile backhaul network with the core/aggregation network, since MPLS TE is always adopted to deploy in the mobile backhaul network for requirements on high reliability, QoS, etc., it has to set up the MPLS TE proxy egress LSP in the mobile backhaul network to stitch with BGP LSP for the end-to-end transport.

This document introduces a new Proxy Destination Object for RSVP-TE. 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. Problem Statement

The typical Seamless MPLS architecture is shown in the figure 1. The typical procedure of setting up the end-to-end transport LSP described in [I-D.ietf-mpls-seamless-mpls] 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 proxy egress LDP LSP with the BGP ingress LSP according to the key of FEC;
3. The remote Aggregation Node (AGN) setup the Ingress LDP LSP to remote Access Node (AN) which has the actual destination.
4. The remote Aggregation Node (AGN) stitches the egress BGP LSP with an ingress LDP 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 network based on Seamless MPLS ([I-D.li-mpls-seamless-mpls-mbb]), 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.

4. 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:

- o Actual destination address: the actual destination address is the destination address of the end-to-end LSP for stitching the proxy egress LSP and the BGP LSP;

- o Proxy destination address: the proxy destination address is the address of Aggregation Node which stitches the proxy egress RSVP-TE LSP and BGP LSP.

When set up the proxy egress RSVP-TE LSP on the Access Node, it must 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. Then 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. At the same time, the Path message carries the actual destination address of the LSP. When the proxy destination node receives the Path message, it sends back the Resv message to allocate label and reserve resource. And the proxy destination node can use the actual destination address to stitch BGP LSP which has the same address as the actual destination address of the proxy egress RSVP-TE LSP.

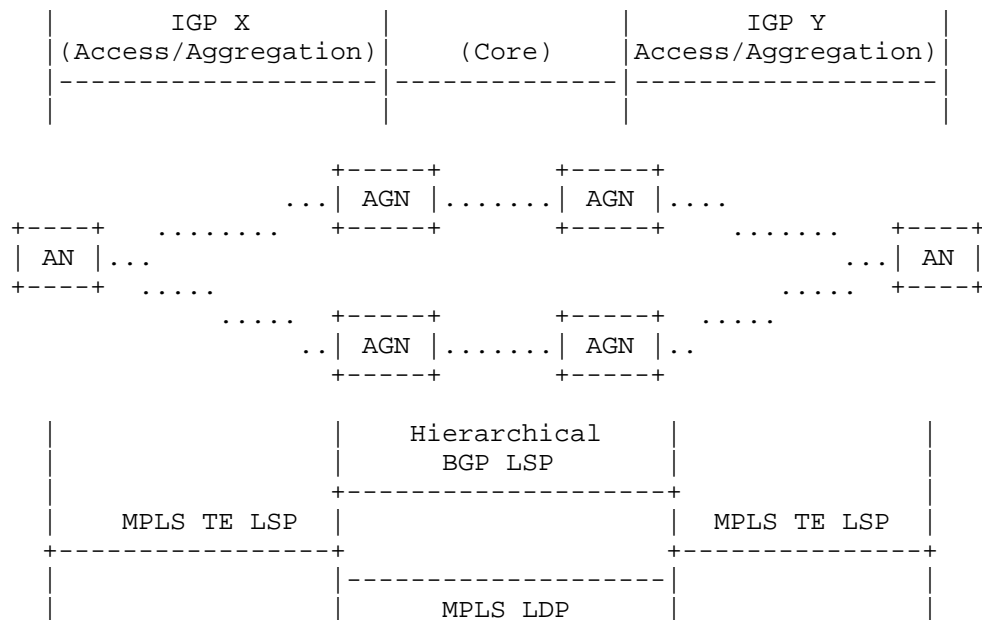


Figure 1 Seamless MPLS Scenario with MPLS TE

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 the actual destination address which 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.

5. Proxy Destination Object

5.1. 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 (TBD)										C-Type (1)											
IPv4 Proxy Destination Address																															

IPv4 Proxy Destination Address: 32 bits. IPv4 address of the proxy destination node of the proxy egress RSVP-TE LSP.

2. IPv6 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 (TBD)										C-Type (2)											
IPv6 Proxy Destination Address																															

IPv6 Proxy Destination Address: 16 bytes. IPv6 address of the proxy destination node 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.

5.2. Procedures

When the ingress LSR 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 and the actual destination address MUST be specified in the Session Object of the Path message. When receive the Resv messages, the ingress LSR SHOULD check if the Proxy Destination object is included. If the Path message includes the Proxy Destination object and the corresponding Resv message does not include this object, the ingress LSR MUST treat the Resv message as wrong messages and MUST NOT set up LSP.

On the transit LSR, 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 LSR to calculate the path, the proxy destination address identified by the Proxy Destination Object MUST be used instead of the actual destination address identified by the Session Object. If the transit LSR 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 messages and MUST NOT set up LSP.

On the egress LSR, when receiving Path messages with Proxy Destination object, it MUST include this object in the corresponding Resv message.

6. IANA Considerations

IANA should allocate Class-Num and C-Type for IPv4 Proxy Destination Object and IPv6 Proxy Destination Object which are defined in this document.

7. Security Considerations

This document does not introduce any additional security issues above those identified in [RFC3209].

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.
- [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.

9.2. Informative 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-07 (work in progress), June 2014.
- [I-D.li-mpls-seamless-mpls-mbb]
Li, Z., Li, L., Morillo, M., and T. Yang, "Seamless MPLS for Mobile Backhaul", draft-li-mpls-seamless-mpls-mbb-01 (work in progress), February 2014.
- [RFC3031] Rosen, E., Viswanathan, A., and R. Callon, "Multiprotocol Label Switching Architecture", RFC 3031, January 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 4, 2015

Zhenbin Li
Lei Li
Huawei Technologies
Manuel Julian Lopez Morillo
Vodafone Group Networks
Tianle Yang
China Mobile
L. Contreras
Telefonica I+D
July 3, 2014

Seamless MPLS for Mobile Backhaul Network
draft-li-mpls-seamless-mpls-mbh-00

Abstract

Seamless MPLS architecture is proposed to integrate access and aggregation/core networks into a single MPLS domain. Through the separation of the service and transport plane Seamless MPLS can provide end to end service independent transport. But when the mobile backhaul network is integrated as the access/aggregation network with the core network based on Seamless MPLS architecture, it will propose new challenges other than the existing solutions for Seamless MPLS. This document introduces the framework of Seamless MPLS to integrate the mobile backhaul network and the core network. The possible challenges of Seamless MPLS for mobile backhaul networks are identified and new requirements are defined accordingly.

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 4, 2015.

Copyright Notice

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

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

Table of Contents

1. Introduction	3
2. Terminology	3
3. Overview of Mobile Backhaul Network	4
4. Scenarios	6
4.1. Scenarios for Network Architecture	6
4.2. Scenarios for Different Edge of Labeled BGP	8
5. Problem Statements and Requirements	11
5.1. Overview	11
5.2. Scalability	12
5.2.1. Problem Statements	12
5.2.2. Requirements	13
5.3. End-to-End Transport	14
5.3.1. Problem Statement	14
5.3.2. Requirements	14
5.4. Hierarchical Service Bearing	15
5.4.1. Problem Statements	15
5.4.2. Requirements	16
5.5. Reliability	16
5.5.1. Problem Statements	16
5.5.2. Requirements	16
5.6. Policy Control	16
5.6.1. Problem Statements	16
5.6.2. Requirements	17
5.7. OAM	17
5.7.1. Problem Statements	17
5.7.2. Requirements	18

6. IANA Considerations	19
7. Security Considerations	19
8. Acknowledgements	19
9. References	19
9.1. Normative References	19
9.2. Informative 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. Through it Seamless MPLS can provide end to end service independent transport. Therefore it removes the need for service specific configurations in network transport nodes.

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 based on Seamless MPLS architecture. Though some approaches of the existing Seamless MPLS architecture 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 introduces the framework of Seamless MPLS to integrate the mobile backhaul network and the aggregation/core network. The possible challenges of Seamless MPLS for mobile backhaul networks are identified and new requirements are defined accordingly. The proposed requirements make it possible to work out the complete solution for a flexible deployment of an end to end mobile backhaul service delivery.

Currently, this document focuses on end to end unicast service deployment. Multicast is out of scope of the document.

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. Overview of Mobile Backhaul Network

The typical mobile backhaul network is shown in the Figure 1. It usually adopts ring topology to save fiber resource and it is always divided into the aggregate network and the access network. In the mobile backhaul network, Cell Site Gateways(CSGs) connect the eNodeBs and RNC Site Gateways(RSGs) connect the RNCs. The mobile traffic is transported from CSGs to RSGs.

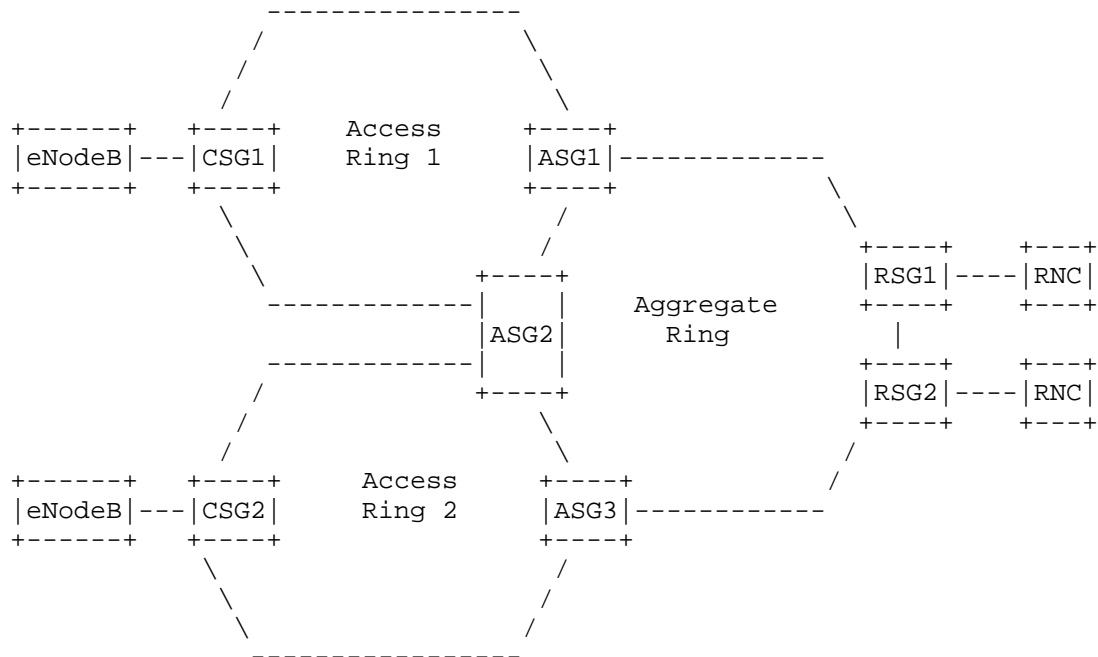


Figure 1 Mobile Backhaul Network

Generally RNCs and eNodeBs connects the local mobile backhaul network. In order to utilize the distributed resource, eNodeBs and RNCs may be located in different networks. That is, the mobile service must span multiple networks including the mobile backhaul networks and the core network. In order to facilitate service provision, Seamless MPLS architecture can be introduced to implement the integration of the mobile backhaul networks and the core network.

[I-D.ietf-mpls-seamless-mpls] defines the possible requirements and solutions for the integration of the access network and the aggregation/core network into a single MPLS domain. 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. Moreover, devices in existing mobile backhaul networks vary in capacity. So there will be different application scenarios and new challenges when Seamless MPLS architecture is applied for the mobile backhaul network.

Note: In order to ease the description, in the following section we will use the PE to represent the CSG in the Seamless MPLS

architecture. In this document the PE and the CSG have the same meaning.

4. 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.

4.1. 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. And the mobile backhaul network including the access network and the aggregation network usually uses a private AS number just for local services. So the integrated end-to-end service means to cross 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.

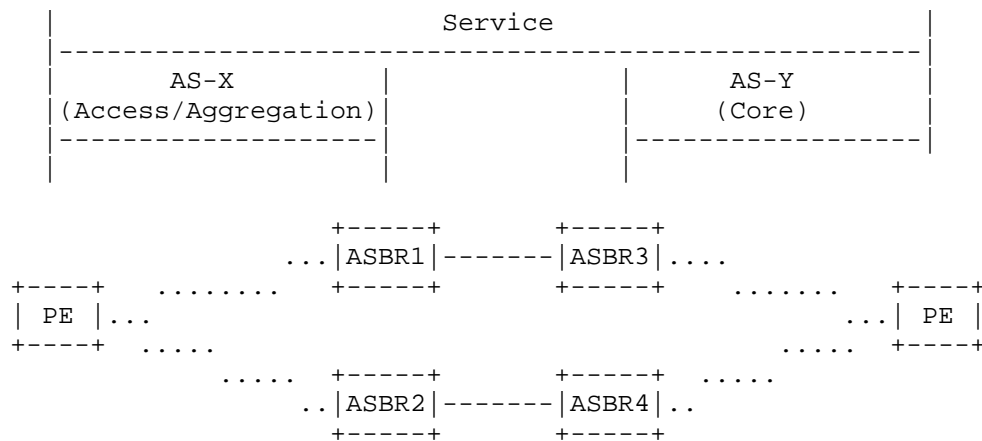


Figure 2 Redundant ASBRs connected Back to Back

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 can distribute the PE's /32 route from ASBRs to ingress PEs in targeted 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.

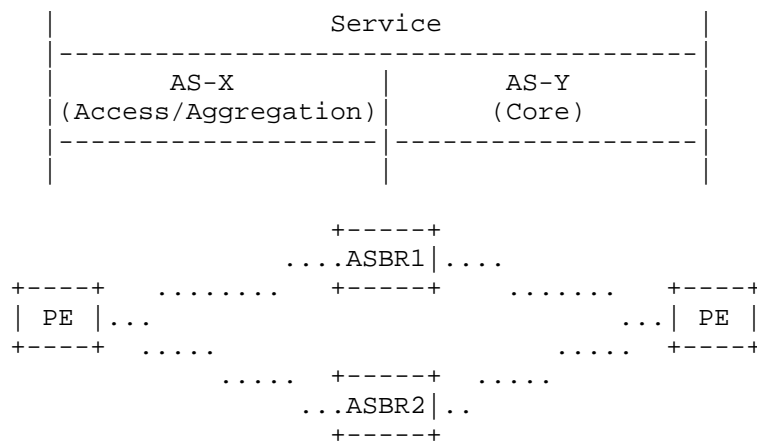


Figure 3 Integrated ASBRs

In this scenario, Seamless MPLS uses label distribution enabled IBGP to establish the end-to-end BGP LSP to support services (that is, it removes the requirement of EBGp sessions) (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 The integrated ASBR should support two different ASes and redistribute the labeled IPv4 routes from one AS to neighboring AS.
- o IBGP can distribute the PE's /32 route from the integrated ASBRs to ingress PEs in targeted 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. In this scenario, the Core network is deployed in a same AS with the access/aggregation network. And the core network, aggregation network and access network are just separated by IGP areas for the reason of scalability.

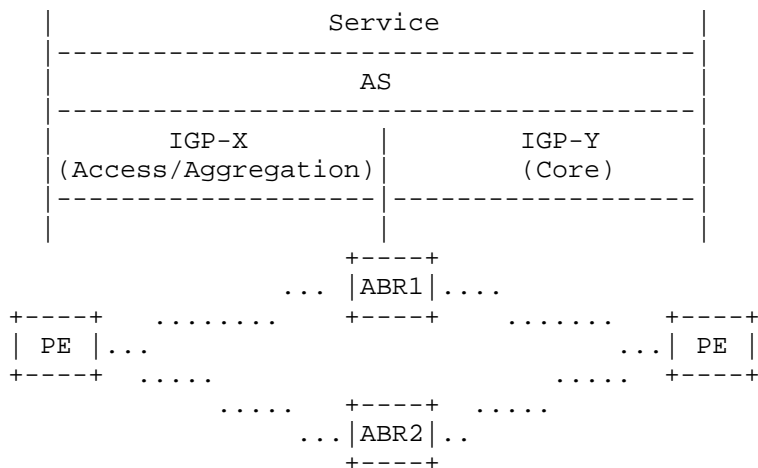


Figure 4 Integrated network in one AS

In this scenario, Seamless MPLS uses labeled IBGP to establish the end-to-end BGP LSP to support services.

4.2. 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 the 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 will 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.

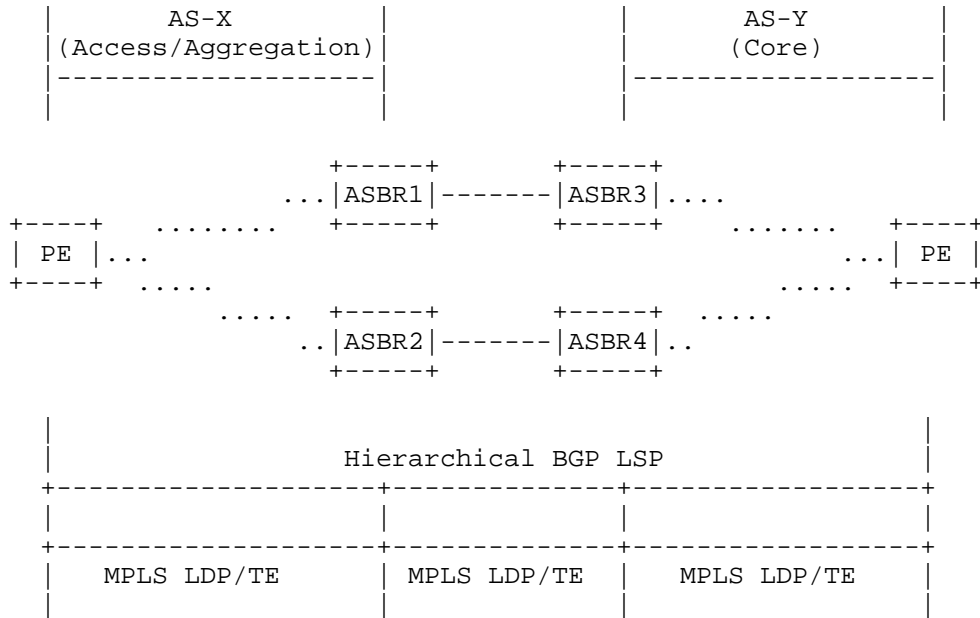


Figure 5 Labeled BGP ended at access devices

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 the access network and BGP LSP in aggregation/core network to support end-to-end services.

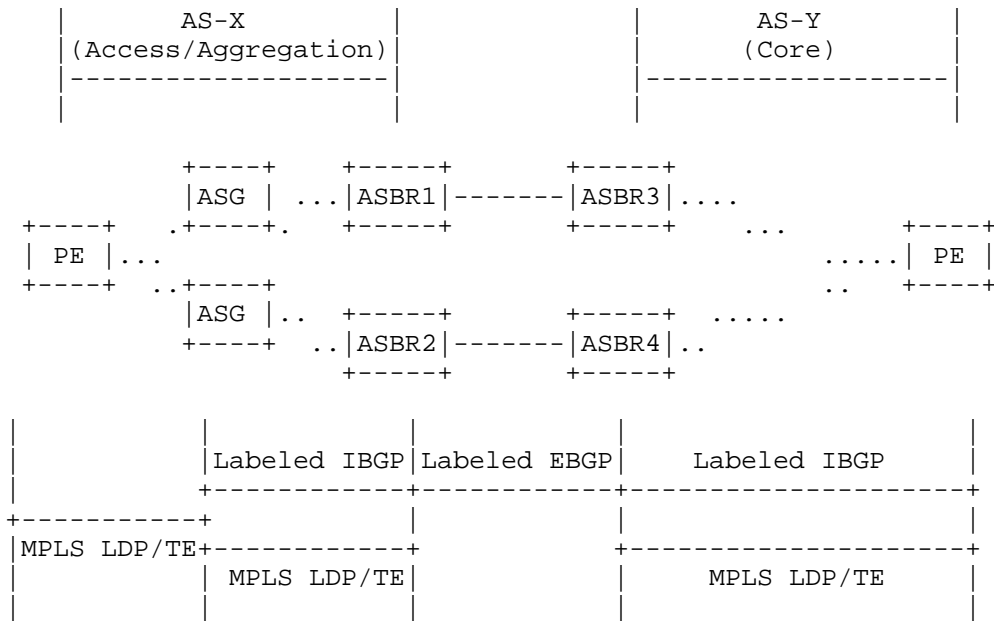


Figure 6 Labeled BGP ended at ASG(P) devices

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.

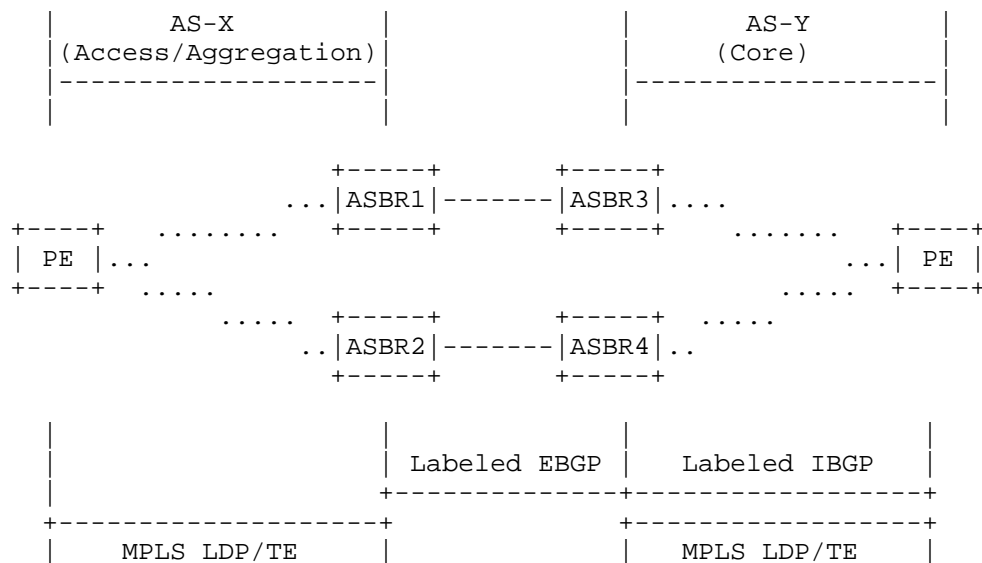


Figure 7 Labeled BGP ended at ASBRs

5. Problem Statements and Requirements

5.1. Overview

Seamless MPLS in [I-D.ietf-mpls-seamless-mpls] describes an architecture by deploying existing protocols like BGP, LDP and ISIS to provides a highly flexible and a scalable architecture and the possibility to integrate 100.000 of nodes. But more requirements for provision of the mobile service and the specialty of the mobile backhaul network proposes new challenges when Seamless MPLS is applied:

1. Challenges from Ring Network

The mobile backhaul network always adopts the ring network deployment. When CSGs access the ring topology and LDP is adopted as the transport plane, there will be multiple hops from CSGs to ASGs/RSGs which has effect on LDP DoD. Moreover the route loop may always happens which will affect the network coverage of the possible IP FRR/LDP FRR solutions such as LFA [RFC6571] for high reliability.

2. Challenges from MPLS TE

In the mobile backhaul network, in order to provide SDH-like service, MPLS TE or MPLS TP technologies are always introduced instead of MPLS LDP for transport of mobile service. When integrate the core network

and the mobile backhaul network, in the transport plane, the interworking between LDP/BGP LSP and RSVP-TE is inevitable. It will have much effect on the end-to-end transport, OAM, protection and scalability.

3. Challenges from 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 L2VPN PW. It needs simplification of flexible policy control and providing complete OAM solutions in different layers.

5.2. Scalability

5.2.1. Problem Statements

There will be huge configuration when MPLS TE is adopted to transport mobile service. 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. In order to satisfy different requirements of the mobile service, it will propose the configuration issues for MPLS TE:

1. Tunnel/LSP Configuration

Since rich set of traffic engineering attributes have to be specified for MPLS TE tunnel and LSP, it has to configure these attributes at the ingress node for each MPLS TE tunnel/LSP to satisfy the requirements such as bandwidth guarantee, fast protection switch, OAM, etc.

2. Path Constraints Configuration

In order to satisfy the requirement of path design for MPLS TE LSP, it has to introduce additional configuration which will deteriorate the configuration work for MPLS TE.

1) Return Path Issue of BFD for MPLS LSPs

In order for high reliability BFD for MPLS LSPs ([RFC5884]) can be used to detect the possible failure fast which can trigger traffic switch between the primary LSP and the backup LSP of a specific MPLS TE tunnel. When BFD for MPLS LSPs is deployed, the return path of BFD traffic may take an IP path which is different from the forward path. The failure that happens in the return path may cause wrong

traffic switch. In order to solve the return path issue of BFD for MPLS LSPs, it has to be guaranteed that the forward path and the return path must be co-routed. For MPLS TE LSPs the explicit path has to be configured for the forward LSP and the return LSP.

2) Completely disjointed primary and backup LSP

MPLS TE Hot-standby feature is always introduced to implement traffic protection. That is, primary LSP and backup LSP are setup at the same time for one MPLS TE tunnel. In order to achieve higher reliability, it is required that the primary and backup LSP should not share any nodes and links. Thus when failure happens in the primary path, the backup LSP can always take over the traffic. So it has to introduce the additional path constraints configuration to satisfy the requirements.

3) Avoid passing through different access rings

When the mobile traffic is transported from the CSG to the RSG, it is expected that the path would not pass through multiple access rings. Since the bandwidth of the access ring is always designed to satisfy its own bandwidth requirement, if mobile traffic from other access ring pass through, the access ring is prone to be overloaded which will cause traffic loss owing to traffic congestion. It is complex to satisfy the requirement using cost, link color or explicit path configurations.

In order to cope with above issues, the complex traffic engineering attributes configuration and path constraints configuration has to be introduced for MPLS TE tunnel/LSP. Calculated using the typical MPLS TE tunnel configuration, there will be hundreds of thousands of command lines need to be configured for MPLS TE. Comparing with LDP, the provision work for MPLS TE is time-consuming and error-prone which has much negative effect on the scalability.

5.2.2. Requirements

When Seamless MPLS is applied to the mobile backhaul network, in order to solve the configuration issue of MPLS TE to improve scalability, following requirements are proposed:

REQ 01: It SHOULD avoid traffic engineering attributes configuration for each MPLS TE tunnel/LSP. Auto tunnel mechanism SHOULD be introduced for provision of MPLS TE tunnel.

REQ 02: It SHOULD simplify the path constraints configuration to cope with the return path issue of BFD for MPLS LSPs.

REQ 03: It SHOULD simplify the path constraints configuration to implement completely disjointed primary and backup LSP.

REQ 04: It SHOULD simplify the path constraints configuration to avoid traffic passing through different access rings.

5.3. End-to-End Transport

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 the access nodes 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 4.2. This section will introduces new challenges and requirements for MPLS TE and LDP to implement stitching solution for the end-to-end transport.

5.3.1. Problem Statement

1. Proxy Egress MPLS TE LSP

In the typical Seamless MPLS architecture, LDP DoD are adopted to setup the segment LSP which is stitched with BGP LSP. When MPLS TE is adopted to deploy in the mobile backhaul network, it is necessary to stitch the MPLS TE LSP set up in the mobile backhaul network and BGP LSP in the core network at the stitching point. Since the actual destination of the MPLS TE LSP may be not located in the local mobile backhaul network, it has to set up the proxy egress MPLS TE LSP from the CSG to the stitching point.

2. Multi-hop LDP DoD

In the typical Seamless MPLS architecture, the static route can be configured for set up LDP DoD LSP. In addition, LDP Extension for Inter-Area Label Switched Paths[RFC5283] can be introduced to reduce the numbers of routes. 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 cannot configure the default route for setup of LDP DoD LSP. If static routes are used, it will be troublesome to configure static routes to a specific destination on all nodes in the mobile backhaul network.

5.3.2. Requirements

REQ 11: It SHOULD set up MPLS TE proxy egress LSP to stitch with BGP LSP to implement end-to-end transport.

REQ 12: It SHOULD simplify the route configuration to setup multi-hop LDP DoD LSP.

5.4. Hierarchical Service Bearing

5.4.1. Problem Statements

Though Seamless MPLS can provide end to end service independent transport and removes the need for service specific configurations in network transport nodes, owing to the limited capability of access nodes it may be necessary to introduce hierarchical MPLS-based service bearing solutions.

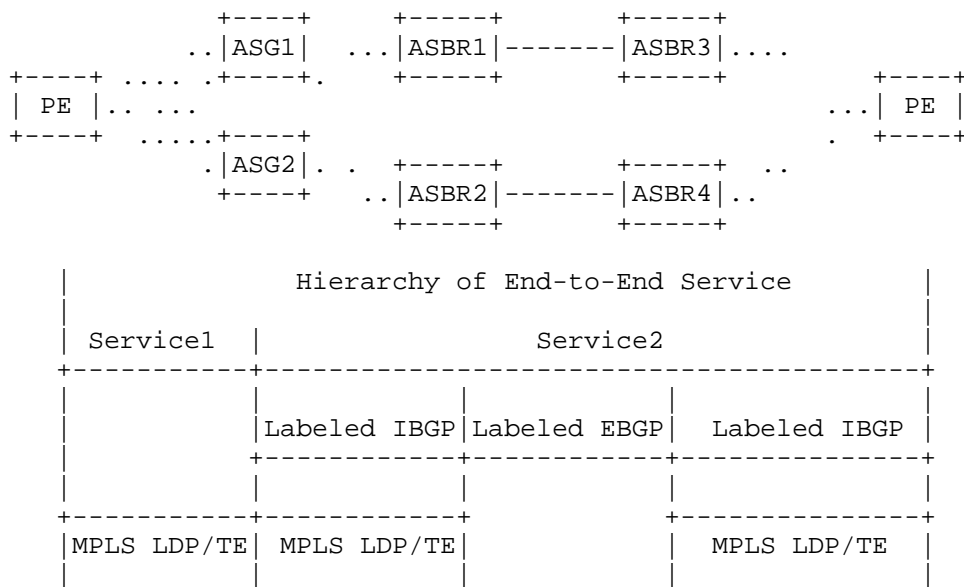


Figure 8 Architecture of Service Layer Stitching

As shown in the above figure, the access nodes (PEs) may be not able to update to support end to end transport, it can utilize the simple hierarchical MPLS-based service bearing solutions such as Hierarchy of VPN (HoVPN) to stitch two segmented VPNs at the ASG or ASBR to establish a VPN service across different domains. The segmented VPN can simply steer the traffic from the access nodes to the ASG or ASBR with higher capability for complex process. This can also simplify the process of access devices (CSGs) and reduce the capability requirement on lower level network devices. Seamless MPLS is not to stop hierarchical service bearing which may need service specific configurations in network transport nodes. On the contrary, it is to provide more flexibility for MPLS-based service bearing. Depending

on the characteristic of the mobile service, different hierarchical service bearing solutions based on L2VPN or L3VPN should be adopted.

5.4.2. Requirements

In order to reduce the requirement on lower level network devices, hierarchical MPLS-based service bearing solutions can be introduced for mobile service. Following requirements are proposed:

REQ 31: Hierarchical L3VPN solutions MAY be introduced to bear L3-based mobile service.

REQ 32: Hierarchical L2VPN solutions MAY be introduced to bear L2-based mobile service.

5.5. Reliability

5.5.1. Problem Statements

The ring topology is always adopted in the mobile backhaul network. The route loop will be bound to happen in the ring network when LFA solution is applied. This means that the backup route does not exist in specific nodes of the ring network according to LFA. Regarding Remote LFA FRR [I-D.ietf-rtgwg-remote-lfa], though it can improve the network coverage comparing with LFA, it still faces the route loop challenges for the back path. In addition, when R-LFA is adopted, there has to set up LDP remote sessions which will propose the scalability issue for fast protection. 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.

5.5.2. Requirements

REQ 41: Scalable IP/LDP FRR solutions SHOULD be provided for the purpose of 100% network coverage when LDP is used for transport of mobile service.

5.6. Policy Control

5.6.1. Problem Statements

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.

5.6.2. Requirements

5.7. 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 management and performance monitoring.

When VPN is used to bear mobile service, multiple layers come into play for implementing the VPN service. This layering extends to the set of OAM protocols that are involved in the ongoing maintenance and diagnostics of VPN networks. The figure below depicts the OAM layering, and shows which devices have visibility into what OAM layer(s).

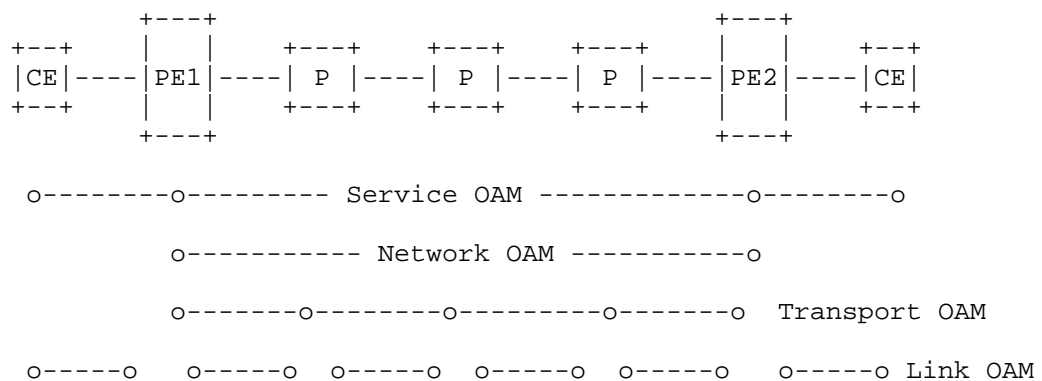


Figure 9 VPN OAM Layering

OAM mechanisms is relatively complete and mature for L2VPN services. However, L3VPN are introduced in the mobile backhaul network for better scalability. The multiple layers for an L3VPN service are as follows:

- The Service Layer runs end to end between the sites that are being interconnected by the L3VPN solution. It is always IP network.
- The Network Layer extends in between the L3VPN PE nodes and is mostly transparent to the core nodes (except where Flow Entropy comes into play). It leverages MPLS for service (i.e. VRF) multiplexing.
- The Transport Layer is dictated by the networking technology of the PSN. It may be either based on MPLS LSPs or IP.
- The Link Layer is dependent upon the physical technology used. Ethernet is a popular choice for this layer, but other alternatives are deployed (e.g. POS, DWDM etc...).

The existing OAM mechanisms for IP and L3VPN is not sufficient to satisfy the OAM requirement of the mobile service, especially for performance monitoring.

2. Flat End-to-End OAM Mechanism

Seamless MPLS provides an architecture to support end-to-end services across multi-separated IP/MPLS domains. Existing path detection 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 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.

5.7.2. Requirements

REQ 61: Performance monitoring mechanism SHOULD be provided for the IP flow.

REQ 62: Performance monitoring mechanism SHOULD be provided for the VPN flow between a pair of L3VPN members.

REQ 63: The end-to-end path trace mechanism SHOULD be provided for the IP flow to collect the whole path information in multiple layers.

6. IANA Considerations

This document makes no request of IANA.

7. Security Considerations

TBD.

8. 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.

9. References

9.1. Normative References

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

9.2. Informative References

- [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-07 (work in progress), June 2014.
- [I-D.ietf-rtgwg-remote-lfa]
Bryant, S., Filsfils, C., Previdi, S., Shand, M., and S. Ning, "Remote LFA FRR", draft-ietf-rtgwg-remote-lfa-06 (work in progress), May 2014.
- [RFC4364] Rosen, E. and Y. Rekhter, "BGP/MPLS IP Virtual Private Networks (VPNs)", RFC 4364, February 2006.
- [RFC5283] Decraene, B., Le Roux, J.L., and I. Minei, "LDP Extension for Inter-Area Label Switched Paths (LSPs)", RFC 5283, July 2008.
- [RFC5884] Aggarwal, R., Kompella, K., Nadeau, T., and G. Swallow, "Bidirectional Forwarding Detection (BFD) for MPLS Label Switched Paths (LSPs)", RFC 5884, June 2010.

[RFC6571] Filsfils, C., Francois, P., Shand, M., Decraene, B., Uttaro, J., Leymann, N., and M. Horneffer, "Loop-Free Alternate (LFA) Applicability in Service Provider (SP) Networks", RFC 6571, June 2012.

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

Luis M. Contreras
Telefonica I+D
Ronda de la Comunicacion, Sur-3 building, 3rd floor
Madrid 28050
Spain

Email: luismiguel.contrerasmurillo@telefonica.com
URI: <http://people.tid.es/LuisM.Contreras/>

Network Working Group
Internet-Draft
Intended status: Informational
Expires: January 4, 2015

Z. Li
Z. Zhuang
J. Dong
Huawei Technologies
July 3, 2014

A Framework for Service-Driven Co-Routed MPLS Traffic Engineering LSPs
draft-li-mpls-serv-driven-co-lsp-fmwk-03

Abstract

MPLS Traffic Engineering (TE) has been widely deployed to satisfy all kinds of requirements of traffic engineering for transport of services. Complexity of configuration has much negative effect on the MPLS TE deployment in the large-scale network. The document identifies the configuration issues for MPLS TE deployment and proposes a new mechanism, the service-driven mechanism, by which the setup of co-routed MPLS Traffic-Engineered Label-Switched Paths (TE LSPs) is triggered by the bidirectional service. Then the document proposes the framework for setting up service-driven co-routed MPLS TE LSP for L2VPN and L3VPN.

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 4, 2015.

Copyright Notice

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

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

Table of Contents

1. Introduction	2
2. Terminology	3
3. Problem Statement	4
3.1. Massive Configuration Issue of TE LSPs	4
3.2. Return Path Issue of BFD for MPLS LSPs	5
3.3. Upgrading Issue of Co-routed Bidirectional LSP	6
4. Framework and Procedures	6
4.1. Service-Driven Co-Routed Unidirectional LSPs for L2VPN	7
4.1.1. Framework	7
4.1.2. Procedures	7
4.2. Service-Driven Co-Routed Unidirectional LSPs for L3VPN	9
4.2.1. Framework	9
4.2.2. Procedures	10
5. IANA Considerations	13
6. Security Considerations	13
7. References	13
7.1. Normative References	13
7.2. Informative References	13
Authors' Addresses	14

1. Introduction

Multiprotocol Label Switching (MPLS) traffic engineering (TE) can satisfy specific traffic engineering attributes [RFC2702]. MPLS TE can effectively schedule, allocate, and use existing network resources to provide bandwidth guarantee and traffic protection for transport of services. MPLS TE is being widely deployed to support packet-based services. Since rich set of traffic engineering attributes have to be specified for each LSP and a great deal of configuration has to be done as the number of MPLS TE LSPs increases, a scalable and simple solution is required to implement TE in a

large-scale network and reduce complexity in operation and management of TE LSPs.

LDP LSP setup is topology-driven which is a scalable way to adapt to the large-scale network. The similar way cannot be used for MPLS TE since the traffic engineering attributes should be specified for MPLS TE LSP which is not necessary for LDP LSP. 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 complex policy applied on it. So it is a natural way to combine setup of MPLS TE LSP with the service it bears. Setup of MPLS TE LSP can be triggered automatically by the service instead of explicitly configuring each MPLS TE LSP and corresponding traffic engineering attributes. We call this method as service-driven comparing to topology-driven. Moreover the service-driven method has much advantage in the process of setting up co-routed TE LSPs. The 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.

This document describes the framework of automatic setup of co-routed MPLS TE LSPs on demand of L2VPN and L3VPN services. The mechanism can facilitate the provisioning of services and the TE LSPs greatly.

2. Terminology

This document uses terminology from the MPLS architecture document [RFC3031], the RSVP-TE protocol specification [RFC3209] which inherits from the RSVP specification [RFC2205] and the Provider Provisioned VPN terminology document [RFC4026].

The document introduces two new concepts by which PEs of VPN can be generally categorized into two types:

- o Active PE: the PE triggers the setup of the LSPs and informs the remote PE;
- o Passive PE: the PE complies with the Active PE's suggestion to set up LSPs.

In this document, the terminology of "tunnel" is identical to the "TE Tunnel" defined in Section 2.1 of [RFC3209], which is uniquely identified by a SESSION object that includes Tunnel end point address, Tunnel ID and Extended Tunnel ID. The terminology "LSP" is identical to the "LSP tunnel" defined in Section 2.1 of [RFC3209], which is uniquely identified by the SESSION object together with

SENDER_TEMPLATE (or FILTER_SPEC) object that consists of LSP ID and Tunnel end point address.

3. Problem Statement

3.1. Massive Configuration Issue of TE LSPs

Deployment MPLS TE in a large-scale networks may require configuration of a potentially large number of TE LSPs.

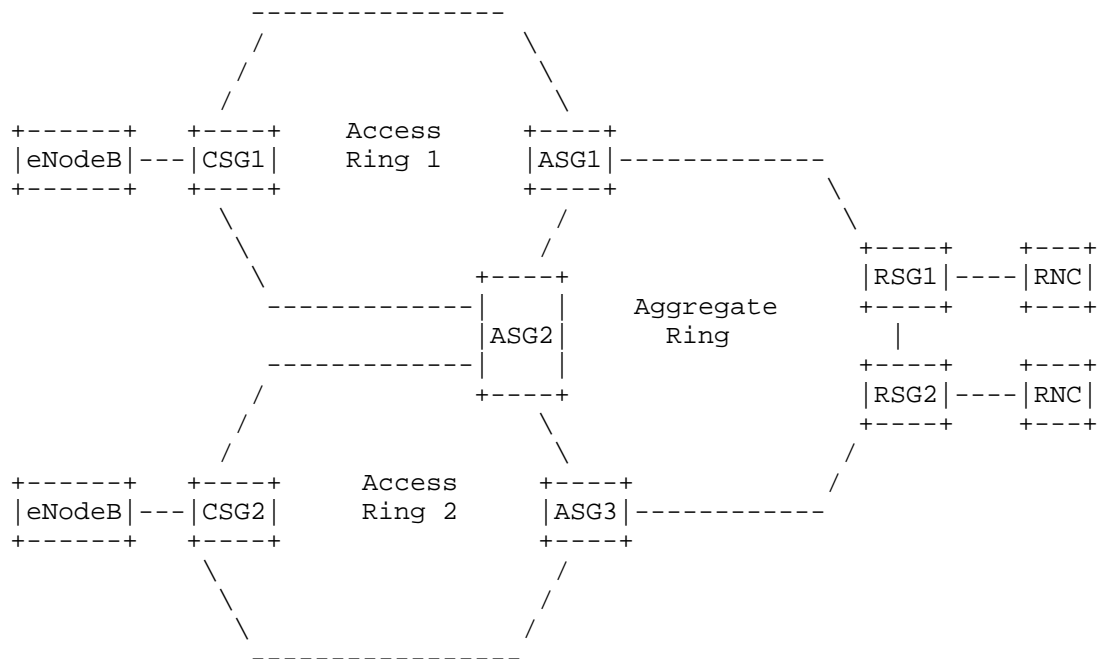


Figure 1 Mobile Backhaul Network

Figure 1 shows an example of the mobile backhaul network. Mobile multimedia devices such as smartphones are ubiquitous now which runs a wide variety of bandwidth-intensive applications and causes unprecedented growth in mobile data traffic. In order to cope with the growth, more cell sites are introduced into the network: more LTE eNodeBs and associated Cell Site Gateways(CSGs) are added in the networks. This causes the network scale expands fast and more and more MPLS TE tunnels need setup between Cell Site Gateways(CSGs) which connect the eNodeBs and RNC Site Gateways(RSGs) which connect the RNCs.

Typically, we assume that:

1. There are 1,000 CSGs need to connect to one RSG.
2. There are three types of bi-directional services transported between one CSG and one RSG. Each type of service needs one VPN and one TE tunnel.
3. There are 10 command lines to configure necessary attributes for each MPLS TE LSP.

So in one RSG it may take 30,000 command lines to set up MPLS TE LSPs. And all CSGs need another 30,000 commands to set up MPLS TE LSPs to one RSG. The huge configuration work is not only time consuming but also prone to mis-configuration. Hence a mechanism to set up MPLS TE LSPs automatically is desirable which can significantly reduce complexity of MPLS TE configuration.

3.2. Return Path Issue of BFD for MPLS LSPs

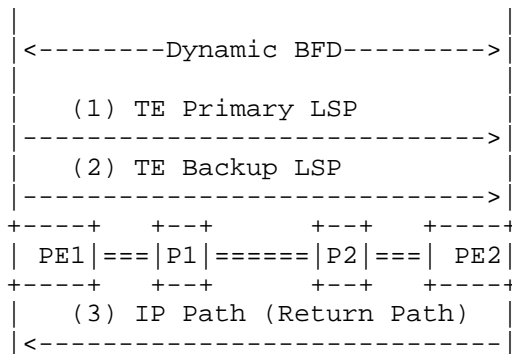


Figure 2: BFD for TE LSPs Scenario

As shown in Figure 2, BFD for MPLS LSPs ([RFC5884]) can be used to detect the possible failure fast which can trigger traffic switch between the primary LSP and the backup LSP. When BFD for MPLS LSPs is deployed, the return path may take an IP path which is different from the forward path. The failure that happens in the return path may cause wrong traffic switch.

In order to solve the return path issue of BFD for MPLS LSPs, it has to be guaranteed that the forward path and the return path must be co-routed. For MPLS TE LSPs the explicit path has to be configured for the forward LSP and the return LSP. In addition, another configuration has to be introduced to bind the return BFD traffic corresponding to the forward BFD traffic to the right return MPLS TE LSP at the ingress node or the egress node. This will deteriorate the configuration work described above. Moreover, if the forward

path changes, the return path may not change accordingly owing to statically binding the forward path and the return path. It will cause that the return path issue of BFD for MPLS LSPs happens again.

3.3. Upgrading Issue of Co-routed Bidirectional LSP

The co-routed bidirectional LSP is defined in [RFC3945]. If co-routed bidirectional LSP is used, the return path is not necessary to configure and the return path issue of BFD for LSPs can be solved naturally. This can simplify operation and management for Service Providers. But it is still necessary to configure each LSP. On the other hand, the unidirectional MPLS TE LSPs have been deployed widely and it is difficult for the service providers to upgrade all possible routers to support co-routed bidirectional LSPs.

4. Framework and Procedures

MPLS TE LSPs depend heavily on manual configuration. So some auto provision methods (e.g. auto mesh [RFC4972]) have been proposed. This document proposes a new mechanism, the service-driven mechanism, to reduce the operation cost of MPLS TE networks.

It is well known that LDP LSP setup is topology-driven which is a scalable way to adapt to the large-scale network. The similar way cannot be used for MPLS TE since the traffic engineering attributes has to be specified for the MPLS TE tunnel. 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 complex policy applied on it. So it is a natural way to trigger MPLS TE LSP setup by the service instead of explicitly configuring each LSP. We call this method as service-driven comparing to topology-driven. In fact BGP-based MVPN ([RFC6513] and [RFC6514]) provides an example of service-driven method which can trigger P2MP TE LSP setup after MVPN membership auto-discovery.

The service-driven method also has much advantage in the process of setting up co-routed MPLS TE LSPs. The 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. This section describes the framework and procedures of setting up the co-routed MPLS TE LSPs. The method needs the signaling of the service advertises the tunnel information between PEs. PEs of VPN can be generally categorized into two types: Active PE and Passive PE. The Passive PE can set up the reverse LSP to the Active PE based on RRO information of the forward LSP which is from the Active PE to the Passive PE. Thus the path of the reverse LSP can be co-routed with the path of the forward LSP.

Service-driven co-routed MPLS TE LSP has following advantages:

- 1) Set up LSPs on demand and save massive configuration effort.
- 2) Reuse existing mechanism as much as possible. It only needs upgrading of PEs instead of whole network upgrading.

4.1. Service-Driven Co-Routed Unidirectional LSPs for L2VPN

4.1.1. Framework

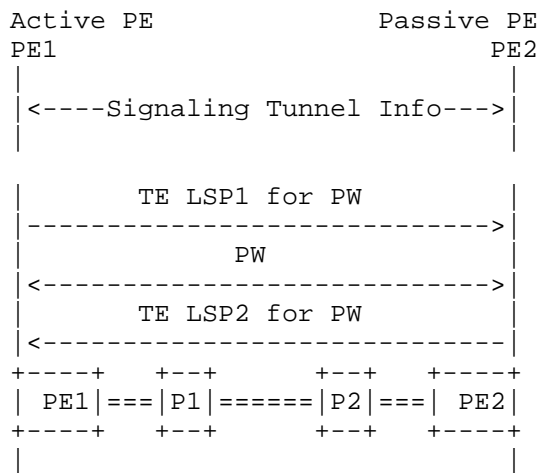


Figure 3: Framework of L2VPN Driven TE LSP

L2VPN, as defined in [RFC4664], is a proven and widely deployed technology. Figure 3 shows a framework for co-routed MPLS TE LSPs driven by L2VPN service. L2VPN is provisioned on PEs and the PW is setup between a pair of PEs. The pair of PEs for a specific PW will be identified as the Active PE and the Passive PE respectively. The Active PE triggers the set up of the forward LSP (TE LSP1) to the Passive PE and advertises the tunnel information to the Passive PE. According to the information advertised by the Active PE, the Passive PE will set up the reverse LSP (TE LSP2) which is co-routed with the forward LSP.

4.1.2. Procedures

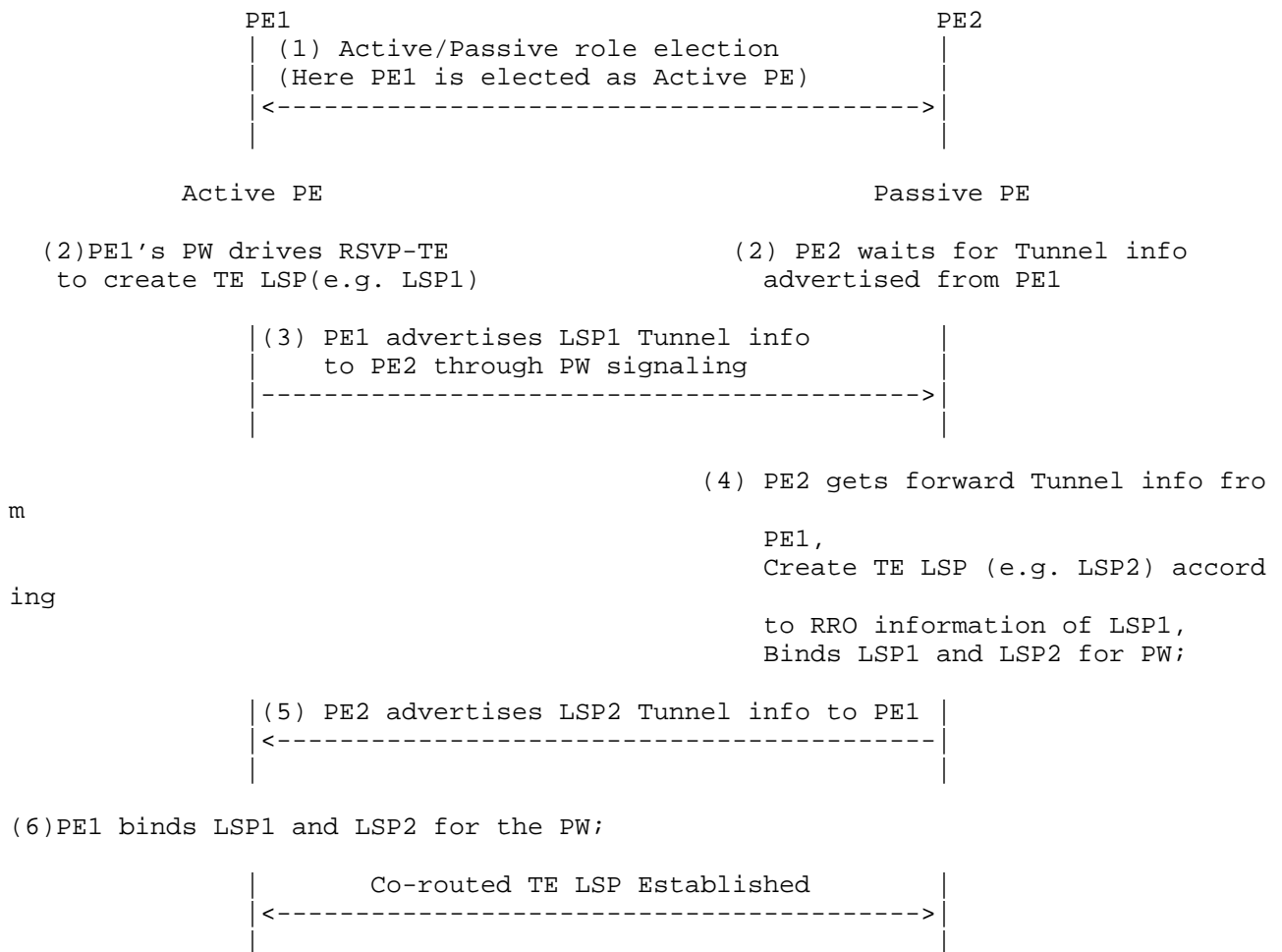


Figure 4: Signaling Procedures of L2VPN Driven Co-Routed TE LSPs

Figure 4 shows the detailed procedures for L2VPN driven co-routed MPLS TE LSPs. Through the above procedures, the co-routed MPLS TE LSPs driven by the PW are established between a pair of PEs.

4.1.2.1. Active/Passive Role Election

The Active and Passive roles of PEs can be determined through manual configuration or dynamic election between a pair of PEs for a specific PW. When the dynamic election method is used, LSR IDs of a pair of PEs between which PW is setup are compared as unsigned integers and the PE with the larger value of LSR ID assumes the Active role.

4.1.2.2. Signaling Tunnel Information

In the service-driven co-routed MPLS TE framework for L2VPN, the tunnel information needs to be advertised between the Active PE and the Passive PE. The Passive PE uses the tunnel information to get corresponding MPLS TE tunnel and RRO information which is used to setup the reverse co-routed MPLS TE LSP.

[I-D.ietf-pwe3-mpls-tp-pw-over-bidir-lsp] defines how the bidirectional Tunnel/LSP identifier information is advertised between a pair of PEs for PW. The similar mechanism can be reused for advertising MPLS TE tunnel/LSP identifier information for service-driven MPLS TE LSPs for L2VPN.

4.1.2.3. Procedures

Step 1: Active/Passive role election through signaling between a pair of PEs of a PW. In this case, assume PE1 as Active PE and PE2 as Passive PE after election;

Step 2: As the active role, the PW service on PE1 drives RSVP-TE to create the forward TE LSP(e.g. LSP1). As the passive role, PE2 waits for tunnel information advertised by PE1;

Step 3: PE1 advertises tunnel information of LSP1 to PE2;

Step 4: PE2 gets tunnel information from PE1 and creates the reverse TE LSP (e.g. LSP2) according to RRO information derived from LSP1. PE2 binds LSP1 and LSP2 for the PW;

Step 5: PE2 advertises tunnel information of LSP2 to PE1;

Step 6: PE1 binds LSP1 and LSP2 for the PW.

Through the above procedures, the co-routed MPLS TE LSPs driven by the PW are established.

4.2. Service-Driven Co-Routed Unidirectional LSPs for L3VPN

4.2.1. Framework

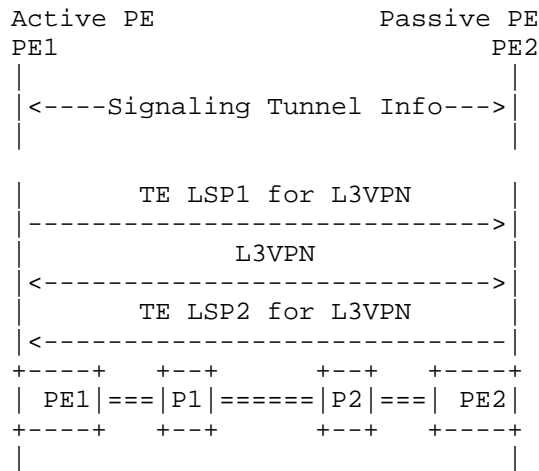


Figure 5: Framework of L3VPN Driven TE LSP

L3VPN services are provided by [RFC4110]. Figure 5 shows a framework for co-routed MPLS TE LSPs driven by L3VPN service. L3VPN is provisioned on PEs and VPN membership is discovered.

The pair of PEs for a specific L3VPN will be identified as the Active PE and the Passive PE respectively. The Active PE triggers the set up of the forward LSP (TE LSP1) to the Passive PE and advertises the tunnel information to the Passive PE. According to the information advertised by the Active PE, the Passive PE will set up the reverse LSP (TE LSP2) which is co-routed with the forward LSP.

4.2.2. Procedures

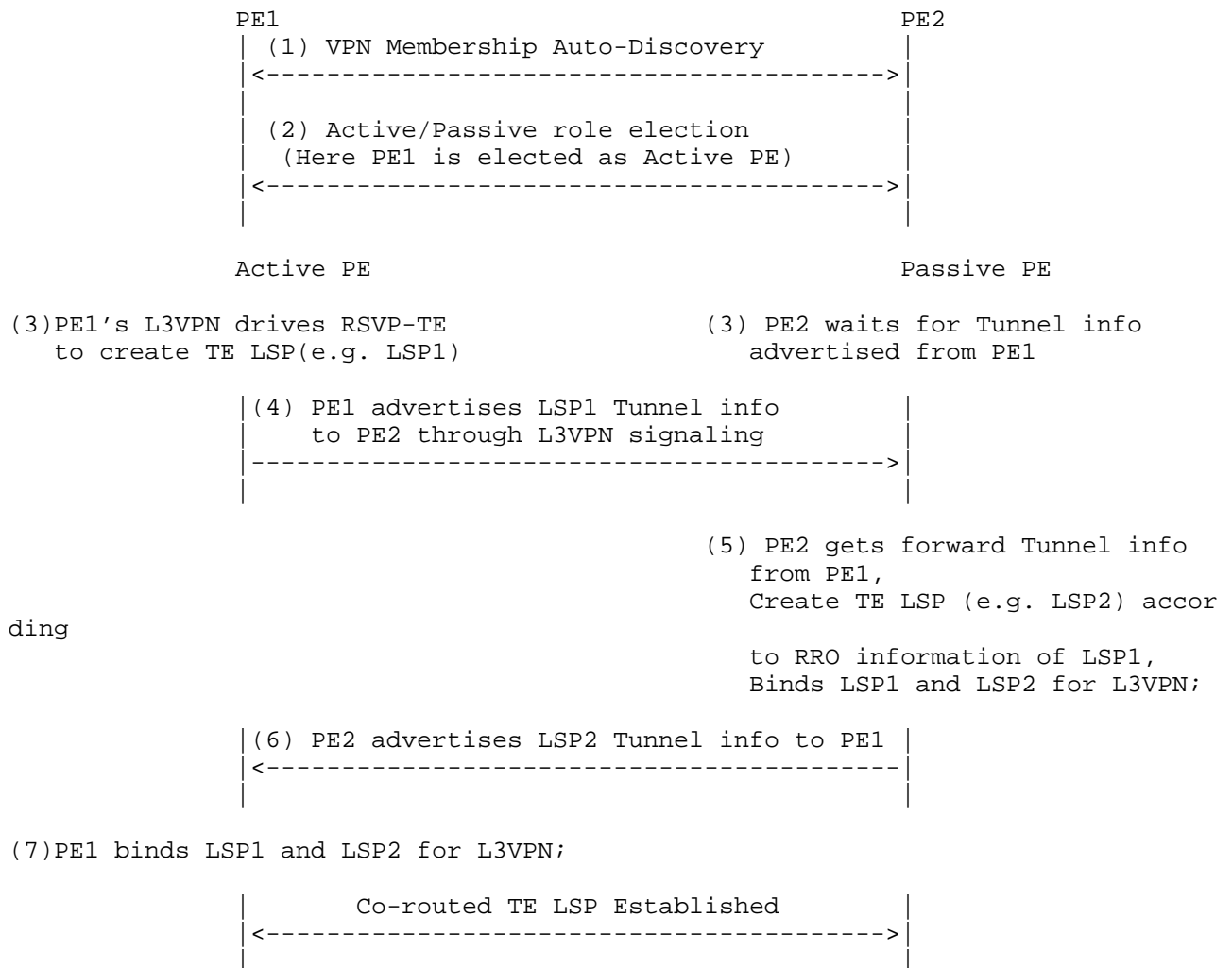


Figure 6: Signaling Procedures of L3VPN Driven Co-Routed TE LSPs

Figure 6 shows the detailed procedures for L3VPN to drive the set up of co-routed MPLS TE LSPs. Through the above procedure, the co-routed MPLS TE LSPs driven by the L3VPN are established between a pair of PEs.

4.2.2.1. VPN Membership Auto-Discovery

In order to set up co-routed MPLS TE LSPs in L3VPN, a point-to-point connection between any two VRFs of a particular VPN needs to be established. VPN membership auto-discovery should be done firstly

and the mechanism defined in [I-D.dong-l3vpn-pm-framework] can be used.

4.2.2.2. Active/Passive Role Election

After obtaining the VPN membership information via VPN membership auto-discovery process, we can identify a pair of VPN members.

The Active and Passive role of PEs can be determined through manual configuration or dynamic election between a pair of PEs for a specific L3VPN. When the dynamic election method is used, LSR IDs of a pair of PEs between which existing the pair of VPN members are compared as unsigned integers and the PE with the larger value of LSR ID assumes the Active role.

4.2.2.3. Signaling Tunnel Information

In the service-driven co-routed MPLS TE framework for L3VPN, the tunnel information needs to be advertised between the Active PE and the Passive PE. The Passive PE uses the tunnel information to get corresponding MPLS TE tunnel and RRO information which is used to setup the reverse co-routed MPLS TE LSP. MP-BGP signaling needs extensions to advertise the MPLS TE tunnel/LSP identifier information for service-driven MPLS TE LSPs for L3VPN.

4.2.2.4. Procedures

Step 1: VPN Membership Auto-Discovery process is done through signaling to identify a pair of VPN members;

Step 2: Active/Passive role election through signaling between a pair of PEs of a L3VPN. In this case, assume PE1 as Active PE and PE2 as Passive PE after election;

Step 3: As the active role, L3VPN service on PE1 drives RSVP-TE to create forward TE LSP(e.g. LSP1), as the passive role, PE2 waits for tunnel information advertised by PE1;

Step 4: PE1 advertises tunnel information of LSP1 to PE2;

Step 5: PE2 gets tunnel information from PE1 and creates TE LSP (e.g. LSP2) according to RRO information derived from LSP1. PE2 binds LSP1 and LSP2 for the L3VPN;

Step 6: PE2 advertises tunnel information of LSP2 to PE1;

Step 7: PE1 binds LSP1 and LSP2 for the L3VPN.

Through the above procedures, the co-routed MPLS TE LSPs driven by the L3VPN are established.

5. IANA Considerations

This document makes no request of IANA.

6. Security Considerations

This document does not change the security properties of L2VPN & L3VPN.

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.
- [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.

7.2. Informative References

- [I-D.dong-l3vpn-pm-framework]
Dong, J., Li, Z., and B. Parise, "A Framework for L3VPN Performance Monitoring", draft-dong-l3vpn-pm-framework-02 (work in progress), January 2014.
- [I-D.ietf-pwe3-mpls-tp-pw-over-bidir-lsp]
Chen, M., Cao, W., Takacs, A., and P. Pan, "LDP extensions for Pseudowire Binding to LSP Tunnels", draft-ietf-pwe3-mpls-tp-pw-over-bidir-lsp-02 (work in progress), November 2013.
- [RFC2205] Braden, B., Zhang, L., Berson, S., Herzog, S., and S. Jamin, "Resource ReSerVation Protocol (RSVP) -- Version 1 Functional Specification", RFC 2205, September 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.

- [RFC3945] Mannie, E., "Generalized Multi-Protocol Label Switching (GMPLS) Architecture", RFC 3945, October 2004.
- [RFC4026] Andersson, L. and T. Madsen, "Provider Provisioned Virtual Private Network (VPN) Terminology", RFC 4026, March 2005.
- [RFC4110] Callon, R. and M. Suzuki, "A Framework for Layer 3 Provider-Provisioned Virtual Private Networks (PPVPNs)", RFC 4110, July 2005.
- [RFC4664] Andersson, L. and E. Rosen, "Framework for Layer 2 Virtual Private Networks (L2VPNs)", RFC 4664, September 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.
- [RFC5884] Aggarwal, R., Kompella, K., Nadeau, T., and G. Swallow, "Bidirectional Forwarding Detection (BFD) for MPLS Label Switched Paths (LSPs)", RFC 5884, June 2010.
- [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

Zhenbin Li
Huawei Technologies
Huawei Bld., No.156 Beiqing Rd.
Beijing 100095
China

Email: lizhenbin@huawei.com

Shunwan Zhuang
Huawei Technologies
Huawei Bld., No.156 Beiqing Rd.
Beijing 100095
China

Email: zhuangshunwan@huawei.com

Jie Dong
Huawei Technologies
Huawei Bld., No.156 Beiqing Rd.
Beijing 100095
China

Email: jie.dong@huawei.com

MPLS Working Group
Internet-Draft
Intended status: Standards Track
Expires: January 1, 2015

G. Mirsky
J. Tantsura
Ericsson
I. Varlashkin
EasyNet
June 30, 2014

Bidirectional Forwarding Detection (BFD) Directed Return Path
draft-mirsky-mpls-bfd-directed-00

Abstract

Bidirectional Forwarding Detection (BFD) is expected to monitor bi-directional paths. When forward direction of a BFD session is to monitor explicitly routed path there is a need to be able to direct far-end BFD peer to use specific path as reverse direction of the BFD session.

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 1, 2015.

Copyright Notice

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

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of

the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction	2
1.1. Conventions used in this document	3
1.1.1. Terminology	3
1.1.2. Requirements Language	3
2. Problem Statement	3
3. Direct Reverse BFD Path	3
3.1. Case of MPLS Data Plane	4
3.1.1. BFD Reverse Path TLV	4
3.1.2. Segment Routing Tunnel sub-TLV	4
3.2. Case of IPv6 Data Plane	5
4. IANA Considerations	6
4.1. TLV	6
4.2. Sub-TLV	6
5. Security Considerations	7
6. Acknowledgements	7
7. Normative References	7
Authors' Addresses	8

1. Introduction

The [RFC5880], [RFC5881], and the [RFC5883] established BFD protocol for IP networks and the [RFC5884] set rules of using BFD Asynchronous mode over IP/MPLS LSPs. All standards implicitly assume that the far-end BFD peer will use the best route regardless of route being used to send BFD control packets towards it. As result, if the near-end BFD peer sends its BFD control packets over explicit path that is diverging from the best route, then reverse direction of the BFD session is likely not to be on co-routed bi-directional path with the forward direction of the BFD session. And because BFD control packets are not guaranteed to cross the same links and nodes in both directions detection of Loss of Continuity (LoC) defect in forward direction is not guaranteed or free of positive negatives.

This document proposes to use BFD Return Path TLV extension to LSP Ping [RFC4379] to instruct the far-end BFD peer to use explicit path for its BFD control packets associated with the particular BFD session. As a special case, forward and reverse directions of the BFD session can form bi-directional co-routed associated channel.

1.1. Conventions used in this document

1.1.1. Terminology

BFD: Bidirectional Forwarding Detection

MPLS: Multiprotocol Label Switching

LSP: Label Switching Path

LoC: Loss of Continuity

1.1.2. Requirements Language

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

2. Problem Statement

BFD is best suited to monitor bi-directional co-routed paths. In most cases, in IP and IP/MPLS networks the best route between two IP nodes is likely to be co-routed in the stable network environment so that implicit BFD requirement is being fulfilled. If BFD is tasked to monitor unidirectional explicitly routed path, e.g. MPLS LSP, its control packets in forward direction would be in-band due to mechanism defined in [RFC5884] and [RFC5586]. But the reverse direction of the BFD session would still follow the best route and that presents following problems in regard to detecting defects on the unidirectional explicit path:

- failure detection on the reverse path cannot be interpreted as bi-directional failure and thus trigger, for example, protection switchover of the forward direction;
- if reverse direction is in Down state, the head-end node would not receive indication of forward direction failure from its far-end peer.

To address these challenges the far-end BFD peer should be instructed to use specific path for its control packets.

3. Direct Reverse BFD Path

3.1. Case of MPLS Data Plane

LSP ping, defined in [RFC4379], uses BFD Discriminator TLV [RFC5884] to bootstrap a BFD session over an MPLS LSP. This document defines a new TLV, BFD Reverse Path TLV, that must contain a single sub-TLV that can be used to carry information about reverse path for the specified in BFD Discriminator TLV session.

3.1.1. BFD Reverse Path TLV

The BFD Reverse Path TLV is an optional TLV within the LSP ping protocol. However, if used the BFD Discriminator TLV MUST be included in an Echo Request message as well. If the BFD Discriminator TLV is not present when the BFD Reverse Path TLV is included, then it MUST be treated as malformed Echo Request, as described in [RFC4379].

The BFD Reverse Path TLV carries the specified path that BFD control packets of the BFD session referenced in the BFD Discriminator TLV are required to follow. The format of the BFD Reverse Path TLV is as presented in Figure 1.

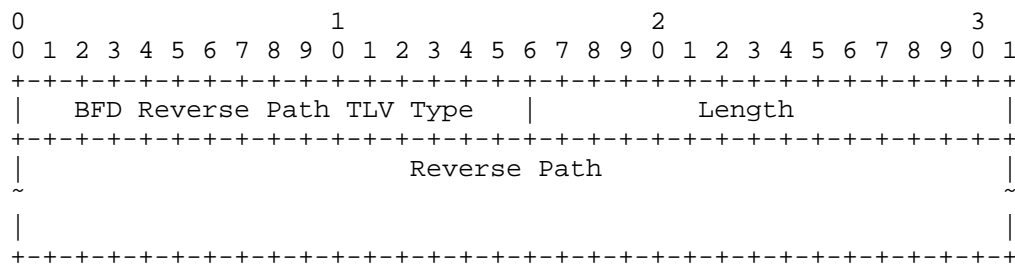


Figure 1: BFD Reverse Path TLV

BFD Reverse Path TLV Type is 2 octets in length and value to be assigned by IANA.

Length is 2 octets in length and defines the length in octets of the Reverse Path field.

3.1.2. Segment Routing Tunnel sub-TLV

With MPLS data plane explicit path can be either Static or RSVP-TE LSP, or Segment Routing tunnel. In case of Static or RSVP-TE LSP [RFC7110] defined sub-TLVs to identify explicit return path. For the Segment Routing with MPLS data plane case a new sub-TLV is defined in this document as presented in Figure 2.

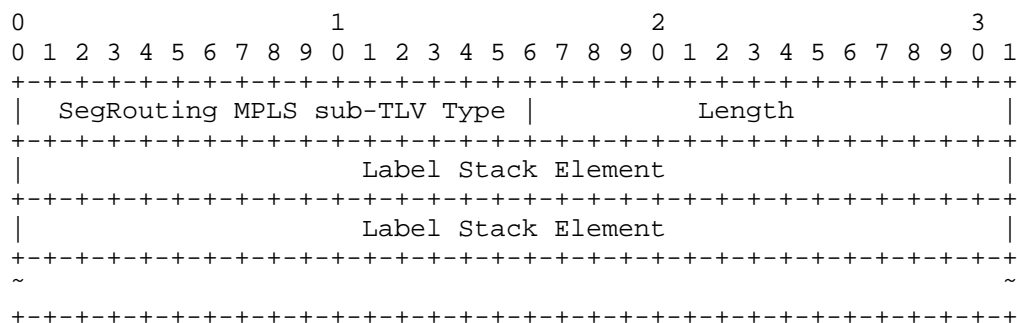


Figure 2: Segment Routing MPLS Tunnel sub-TLV

The Segment Routing Tunnel sub-TLV Type is two octets in length, and will be allocated by IANA.

The Segment Routing Tunnel sub-TLV MAY be used in Reply Path TLV defined in [RFC7110]

3.2. Case of IPv6 Data Plane

IPv6 can be data plane of choice for Segment Routed tunnels [I-D.previdi-6man-segment-routing-header]. In such networks the BFD Reverse Path TLV described in Section 3.1.1 can be used as well. IP networks, unlike IP/MPLS, do not require use of LSP ping with BFD Discriminator TLV[RFC4379] to bootstrap BFD session. But to specify reverse path of a BFD session in IPv6 environment the BFD Discriminator TLV MUST be used along with the BFD Reverse Path TLV. The BFD Reverse Path TLV in IPv6 network MUST include sub-TLV.

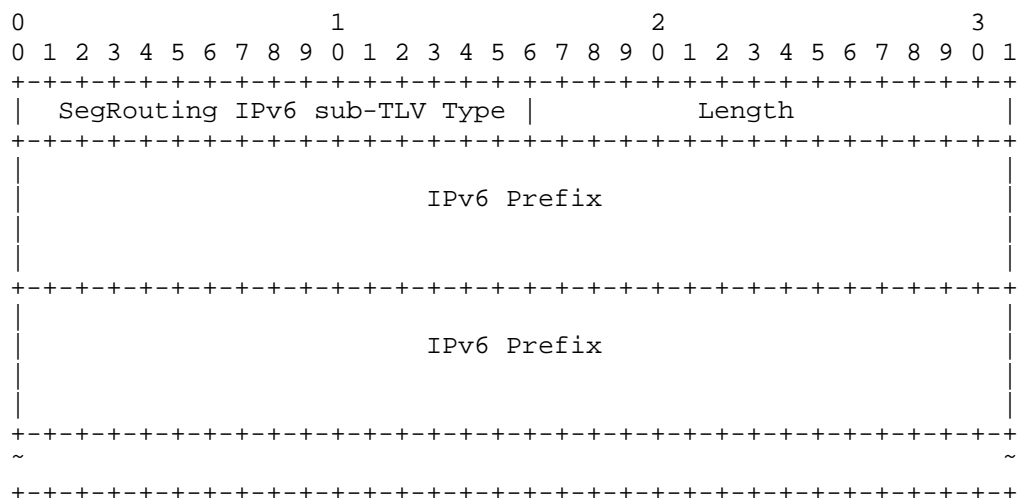


Figure 3: Segment Routing IPv6 Tunnel sub-TLV

4. IANA Considerations

4.1. TLV

The IANA is requested to assign a new value for BFD Reverse Path TLV from the "Multiprotocol Label Switching Architecture (MPLS) Label Switched Paths (LSPs) Ping Parameters - TLVs" registry, "TLVs and sub-TLVs" sub-registry.

Value	Description	Reference
X (TBD1)	BFD Reverse Path TLV	This document

Table 1: New BFD Reverse Type TLV

4.2. Sub-TLV

The IANA is requested to assign one new sub-TLV type from "Multiprotocol Label Switching Architecture (MPLS) Label Switched Paths (LSPs) Ping Parameters - TLVs" registry, "Sub-TLVs for TLV Type 1" sub-registry.

Value	Description	Reference
X (TBD2)	Segment Routing MPLS Tunnel sub-TLV	This document
X (TBD3)	Segment Routing IPv6 Tunnel sub-TLV	This document

Table 2: New Segment Routing Tunnel sub-TLV

5. Security Considerations

Security considerations discussed in [RFC5880], [RFC5884], and [RFC4379], apply to this document.

6. Acknowledgements

7. Normative References

- [I-D.previdi-6man-segment-routing-header]
Previdi, S., Filsfils, C., Field, B., and I. Leung, "IPv6 Segment Routing Header (SRH)", draft-previdi-6man-segment-routing-header-01 (work in progress), June 2014.
- [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.
- [RFC5880] Katz, D. and D. Ward, "Bidirectional Forwarding Detection (BFD)", RFC 5880, June 2010.
- [RFC5881] Katz, D. and D. Ward, "Bidirectional Forwarding Detection (BFD) for IPv4 and IPv6 (Single Hop)", RFC 5881, June 2010.
- [RFC5883] Katz, D. and D. Ward, "Bidirectional Forwarding Detection (BFD) for Multihop Paths", RFC 5883, June 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.

[RFC7110] Chen, M., Cao, W., Ning, S., Jounay, F., and S. Delord,
"Return Path Specified Label Switched Path (LSP) Ping",
RFC 7110, January 2014.

Authors' Addresses

Greg Mirsky
Ericsson

Email: gregory.mirsky@ericsson.com

Jeff Tantsura
Ericsson

Email: jeff.tantsura@ericsson.com

Ilya Varlashkin
EasyNet

Email: Ilya.Varlashkin@easynet.com

MPLS Working Group
Internet-Draft
Intended status: Standards Track
Expires: January 1, 2015

G. Mirsky
S. Ruffini
Ericsson
J. Drake
Juniper Networks
S. Bryant
Cisco Systems
A. Vainshtein
ECI Telecom
June 30, 2014

Residence Time Measurement in MPLS network
draft-mirsky-mpls-residence-time-02

Abstract

This document specifies G-ACh based Residence Time Measurement and how it can be used by time synchronization protocols being transported over MPLS domain.

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 1, 2015.

Copyright Notice

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

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect

to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction	2
1.1. Conventions used in this document	2
1.1.1. Terminology	2
1.1.2. Requirements Language	3
2. Residence Time Measurement	3
3. G-ACh for Residence Time Measurement	4
4. Control Plane Theory of Operation	5
4.1. RSVP-TE Control Plane Operation to Support RTM	5
5. Data Plane Theory of Operation	6
6. Applicable PTP Scenarios	6
7. IANA Considerations	7
7.1. New RTM G-ACh	7
7.2. New RTM TLV Registry	7
8. Security Considerations	7
9. Acknowledgements	8
10. References	8
10.1. Normative References	8
10.2. Informative References	9
Authors' Addresses	9

1. Introduction

Time synchronization protocols, Network Time Protocol version 4 (NTPv4) [RFC5905] and Precision Time Protocol (PTP) Version 2, a.k.a. IEEE-1588 v.2, can be used to synchronized clocks across network domain. In some scenarios calculation of time packet of time synchronization protocol spends within a node, called Residence Time, can improve accuracy of clock synchronization. This document defines new Generalized Associated Channel (G-ACh) that can be used in Multi-Protocol Label Switching (MPLS) network to measure Residence Time over Label Switched Path (LSP). Transport of packets of a time synchronization protocol over MPLS domain is outside of scope of this document.

1.1. Conventions used in this document

1.1.1. Terminology

MPLS: Multi-Protocol Label Switching

ACH: Associated Channel

TTL: Time-to-Live

G-ACh: Generic Associated Channel

GAL: Generic Associated Channel Label

NTP: Network Time Protocol

ppm: part per million

PTP: Precision Time Protocol

LSP: Label Switched Path

LSR: Label Switched Router

OAM: Operations, Administration, and Maintenance

RTM: Residence Time Measurement

IGP: Internal Gateway Protocol

1.1.2. Requirements Language

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

2. Residence Time Measurement

Packet Loss and Delay Measurement for MPLS Networks [RFC6374] can be used to measure one-way or two-way end-to-end propagation delay over LSP or PW. But none of these metrics is useful for time synchronization across a network. For example, PTPv2 uses "residence time", time it takes for a PTPv2 event packet to transit a node. The residence times are accumulated in the correctionField of the PTP event messages or of the associated follow-up messages (or Delay_Resp message associated with the Delay_Req message) in case of two-step clocks. The residence time values are specific to each output PTP port and message.

Note the delay of propagation over a link connected to a port receiving the PTP event message is handled by IEEE 1588 [IEEE.1588.2008] by means of specific messages, Pdelay_Req and Pdelay_Resp, or Delay_Req and Delay_Resp depending on the applicable delay mechanism, peer-to-peer or delay request-response mechanism respectively.

This document proposes mechanism to accumulate packet residence time from all LSRs that support the mechanism across the particular LSP.

3. G-ACh for Residence Time Measurement

RFC 5586 [RFC5586] and RFC 6423 [RFC6423] extended applicability of PW Associated Channel (ACH) [RFC5085] to LSPs. G-ACh presents mechanism to transport OAM and other control messages and trigger their processing by arbitrary transient LSRs through controlled use of Time-to-Live (TTL) value.

Packet format for Residence Time Measurement (RTM) presented in Figure 1

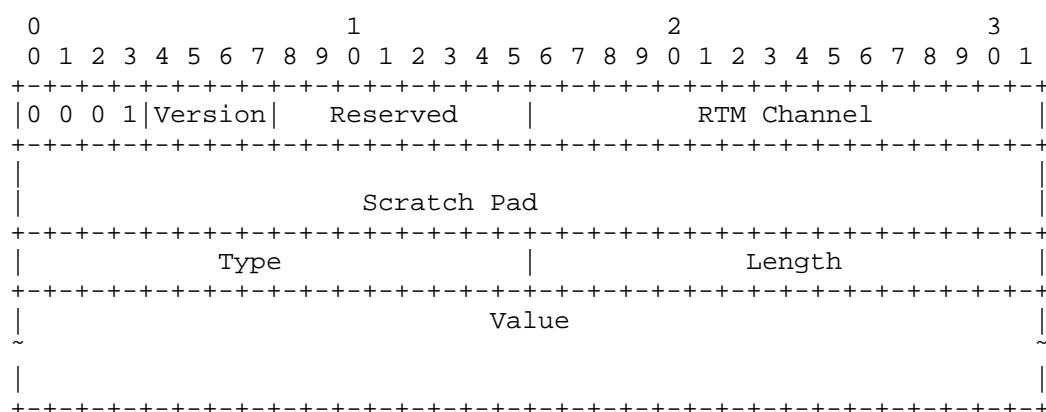


Figure 1: G-ACh packet format for Residence Time Measurement

The Version field is set to 0, as defined in RFC 4385 [RFC4385]. The Reserved field must be set to 0 on transmit and ignored on receipt. The RTM G-ACh field, value to be allocated by IANA, identifies the packet as such. The Scratch Pad field is 8 octets in length and is used to accumulate the residence time spent in LSRs transited by the packet on its path from ingress LSR to egress LSR. Its format is IEEE double precision and its units are nanoseconds.

The Type field identifies type of Value that the TLV carries. IANA will be asked to create sub-registry in Generic Associated Channel (G-ACh) Parameters Registry called "MPLS RTM TLV Registry". The Length field is number of octets of the Value field. The optional Value field may be used to carry a packet of a given time synchronization protocol. If the packet carried in the RTM message, then it accordingly identified by distinct Type, and may be NTP [RFC5905] or PTP [IEEE.1588.2008]. It is important to note that the packet may be authenticated or encrypted and carried over MPLS LSP

edge to edge unchanged while residence time being accumulated in the Scratch Pad field. The TLV MUST be included in the RTM.

4. Control Plane Theory of Operation

A router will announce its support for RTM in a new sub-TLV, the RTM Capable TLV which will be defined in a subsequent version of this document, for the router capabilities TLV defined in RFC 4970 (OSPF) [RFC4970] and RFC 4971 (IS-IS) [RFC4971].

The operation of RTM depends upon TTL expiry to deliver an RTM packet from one RTM capable LSR to the next along the path from ingress LSR to egress LSR, which means that an RTM capable LSR needs to be able to compute a TTL which will cause the expiry of an RTM packet on the next RTM capable LSR.

However, because of Equal Cost Multipath, labels distributed by LDP do not instantiate a single path between a given ingress/egress LSR pair but rather a graph and different flows will take different paths through this graph. This means one doesn't know the path that RTM packets will take or even if they all take the same path. So, in an environment in which not all routers in an IGP domain support RTM, it is effectively impossible to use TTL expiry to deliver RTM packets and hence RTM cannot be used for LSPs instantiated using LDP. In the special but important case of environment in which all routers in an IGP domain support RTM, setting the TTL to 1 will always cause the expiry of an RTM packet on the next RTM capable downstream LSR and hence in such an environment, RTM can be used for LSPs instantiated using LDP.

Generally speaking, RTM is more useful for an LSP instantiated using RSVP-TE [RFC3209] because the LSP's path can be known.

4.1. RSVP-TE Control Plane Operation to Support RTM

An ingress LSR that wishes to perform RTM along a path through an MPLS network to an egress LSR verifies that the selected egress LSR supports RTM via the egress LSR's advertisement of the RTM Capable TLV. In the Path message that the ingress LSR uses to instantiate the LSP to that egress LSR it places initialized Record Route and RTM Set (see below) Objects, which tell the egress LSR that RTM is desired for this LSP.

In the Resv message that the egress LSR sends in response to the received Path message, it includes initialized Record Route and RTM Set objects. The latter object will be defined in a subsequent version of this document and it contains an ordered list, from egress LSR to ingress LSR, of the RTM capable LSRs along the LSP's path.

Each such LSR will use the ID of the first LSR in the RTM Set Object in conjunction with the Record Route Object to compute the hop count to its downstream RTM capable LSR. It will also insert its ID at the beginning of the RTM Set Object before forwarding the Resv upstream.

After the ingress LSR receives the Resv, it will begin sending RTM packets to the first RTM capable LSR on the LSP's path. Each RTM packet has its Scratch Pad field initialized and its TTL set to expire on that LSR.

It should be noted that RTM can also be used for LSPs instantiated using [RFC3209] in an environment in which all routers in an IGP support RTM. In this case the RTM Set Object is not used.

5. Data Plane Theory of Operation

After instantiating an LSP for a path using RSVP-TE [RFC3209] as described in Section 4.1 or if this is the special case of homogeneous RTM-capable IP/MPLS domain discussed in the last paragraph of Section 4, ingress LSR MAY begin sending RTM packets to the first RTM capable downstream LSR on that path. Each RTM packet has its Scratch Pad field initialized and its TTL set to expire on the next downstream LSR. Each RTM capable LSR that receives an RTM packet records the time at which it receives that packet as well as the time at which it transmits that packet; this should be done as close to the physical layer as possible. Just prior to sending that packet, it takes the difference between those two times and adds it to the value in the Scratch Pad field. Note, for the purpose of calculating a residence time, a free running clock may be sufficient, as, for example, 4.6 ppm accuracy leads to 4,6 ns error for residence time in the order of 1 ms.

The RTM capable LSR also sets the RTM packet's TTL to expire on the next RTM capable downstream from it LSR.

The egress LSR may then use the value in the Scratch Pad field to perform time correction. For example, the egress LSR may be a PTP Boundary Clock synchronized to a Master Clock and will use the value in the Scratch Pad Field to update PTP's Correction Field.

6. Applicable PTP Scenarios

The proposed approach can be directly integrated in a PTP network based on delay request-response mechanism. The RTM capable LSR nodes act as end-to-end transparent clocks, and typically boundary clocks, at the edges of the MPLS network, use the value in the Scratch Pad

field to update the correctionField of the corresponding PTP event packet prior to performing the usual PTP processing.

Under certain assumptions the proposed solution in a network where peer delay mechanism is used is also possible. The solution in this case requires the definition of a specific protocol to be used to calculate the link delays according to a peer delay link measurement approach. This is not described in this version of the draft.

7. IANA Considerations

7.1. New RTM G-ACh

IANA is requested to reserve a new G-ACh as follows:

Value	Description	Reference
X	Residence Time Measurement	This document

Table 1: New Residence Time Measurement

7.2. New RTM TLV Registry

IANA is requested to create sub-registry in Generic Associated Channel (G-ACh) Parameters Registry called "MPLS RTM TLV Registry". All code points within this registry shall be allocated according to the "IETF Review" procedure as specified in [RFC5226] This document defines the following new values RTM TLV type

Value	Description	Reference
0	Reserved	This document
TBD1	No payload	This document
TBD2	PTPv2	This document
TBD3	NTP	This document

Table 2: RTM TLV Type

8. Security Considerations

Routers that support Residence Time Measurement are subject to the same security considerations as defined in [RFC5586] and [RFC6423].

9. Acknowledgements

TBD

10. References

10.1. Normative References

- [IEEE.1588.2008] "Standard for a Precision Clock Synchronization Protocol for Networked Measurement and Control Systems", IEEE Standard 1588, March 2008.
- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 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.
- [RFC4385] Bryant, S., Swallow, G., Martini, L., and D. McPherson, "Pseudowire Emulation Edge-to-Edge (PWE3) Control Word for Use over an MPLS PSN", RFC 4385, February 2006.
- [RFC4970] Lindem, A., Shen, N., Vasseur, JP., Aggarwal, R., and S. Shaffer, "Extensions to OSPF for Advertising Optional Router Capabilities", RFC 4970, July 2007.
- [RFC4971] Vasseur, JP., Shen, N., and R. Aggarwal, "Intermediate System to Intermediate System (IS-IS) Extensions for Advertising Router Information", RFC 4971, July 2007.
- [RFC5085] Nadeau, T. and C. Pignataro, "Pseudowire Virtual Circuit Connectivity Verification (VCCV): A Control Channel for Pseudowires", RFC 5085, December 2007.
- [RFC5586] Bocci, M., Vigoureux, M., and S. Bryant, "MPLS Generic Associated Channel", RFC 5586, June 2009.
- [RFC5905] Mills, D., Martin, J., Burbank, J., and W. Kasch, "Network Time Protocol Version 4: Protocol and Algorithms Specification", RFC 5905, June 2010.
- [RFC6423] Li, H., Martini, L., He, J., and F. Huang, "Using the Generic Associated Channel Label for Pseudowire in the MPLS Transport Profile (MPLS-TP)", RFC 6423, November 2011.

10.2. Informative References

- [RFC5226] Narten, T. and H. Alvestrand, "Guidelines for Writing an IANA Considerations Section in RFCs", BCP 26, RFC 5226, May 2008.
- [RFC6374] Frost, D. and S. Bryant, "Packet Loss and Delay Measurement for MPLS Networks", RFC 6374, September 2011.

Authors' Addresses

Greg Mirsky
Ericsson

Email: gregory.mirsky@ericsson.com

Stefano Ruffini
Ericsson

Email: stefano.ruffini@ericsson.com

John Drake
Juniper Networks

Email: jdrake@juniper.net

Stewart Bryant
Cisco Systems

Email: stbryant@cisco.com

Alexander Vainshtein
ECI Telecom

Email: Alexander.Vainshtein@ecitele.com

MPLS Working Group
INTERNET-DRAFT
Intended Status: Proposed Standard
Expires: January 3, 2015

R. Singh
Y. Shen
J. Drake
Juniper Networks
July 2, 2014

Entropy label for seamless MPLS
draft-ravisingh-mpls-el-for-seamless-mpls-02

Abstract

This document describes certain optimizations to how entropy labels can be used for load balancing in a seamless MPLS architecture, as enabled by LSP concatenation and LSP hierarchies.

The definition of the control plane and data plane behavior at LSP concatenation points; and at the ingress of an LSP in a hierarchy of LSPs, as described in this document, brings the benefits of entropy labels to certain deployment scenarios that may not have had such benefits as specified in [EL-RFC].

This document, if approved, updates RFC 6790.

Status of this Memo

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

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

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

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

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

This Internet-Draft will expire on January 3, 2015.

Copyright and License 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.	Terminology	5
3.	Key attributes of the entropy label solution: Summary from [EL-RFC]	6
4.	Problems and Motivation	7
4.1	EL applicability for seamless MPLS	7
4.2	EL for LSP concatenation	8
4.2.1	Spectrum of EL usage behaviors required to be supported for concatenated LSPs	9
4.2.1.1	Entropy label for per-segment LSP	9
4.2.1.2	Entropy label for notional-segment-LSP(s)	10
4.2.1.3	Entropy label for e2e LSP	10
4.3	EL for LSP hierarchy	10
4.3.1	Possibility of unnecessary reduction of max-payload of the LSP:	11
4.3.2	Possibility of EL being non-usable for load-balancing:	12
5.	EL for LSP concatenation/hierarchy	13
5.1	Additional EL abstractions: specific to LSP concatenation/hierarchy	13
5.2	New abstractions: EL applicability for LSP concatenation	13
5.2.1	Signaling	13
5.2.1.1	Signaling ELC at concatenation points (Translation rules)	15
5.2.2	Data plane aspects	16
5.2.2.1	LSP concatenation: Differing EL dispositions	16
5.3	New abstractions: EL applicability for LSP hierarchy	18
5.3.1	Management plane:	18
5.3.2	Data plane aspects	19
6.	Use-cases	20

6.1 Carrier of carrier L3VPN	21
6.2 Inter-AS L3VPN: Option C	22
7. Manageability	22
8. Security considerations	22
9. Acknowledgments	22
10. IANA considerations	22
11. References	23
11.1 Normative References	23
11.2 Informative References	23
Authors' Addresses	24

1 Introduction

[EL-RFC] specifies a way to implement load-balancing in an MPLS network such that sub-flows of an LSP may be identified and sent on different paths through the network. This is achieved by using entropy labels (ELs) to abstract out the flow-identifying information into the entropy label and inserting the entropy label underneath the LSP label. The transit LSRs perform the load-balancing hash-computation, on the label-stack alone, to effect a good load-balancing outcome without a need to parse inner headers.

The key feature of [EL-RFC] is that it defines the EL in the context of a given LSP. [EL-RFC] defines the signaling extensions by which entropy label capability might be signaled for LSPs setup by RSVP-TE, LDP or [LU-BGP]. While that works well for individual LSPs, there are additional issues to consider for the seamless MPLS architecture [S-MPLS].

The currently-under-definition framework for seamless MPLS proposes an architecture ([S-MPLS]) that shall enable the setting-up of MPLS LSPs from access nodes to access nodes using a medley of signaling protocols and statically configured LSPs...by essentially leveraging LSP concatenation and hierarchies to carry labeled traffic in larger portions of the network without essentially increasing control plane state. There are special EL-related considerations that need to be dealt with to make EL more suitable for seamless MPLS, on account of its reliance on LSP concatenation and hierarchies.

This document defines additional abstractions and rules for the use of entropy-label with LSP concatenation/hierarchy to enable the use of ELs for the seamless MPLS architecture. This document describes how entropy labels may be used when the LSP has been setup by concatenation LSP segments or by tunneling the LSP over other LSPs. It is conceivable that different signaling protocols are in use to create an e2e LSP.

LSP stitching is the process of connecting LSP segments in the data plane to form a single e2e data plane LSP. This is achieved by setting up LSP segments through signaling or through management action, and then setting-up an e2e LSP that "uses" these LSP segments as hops in its path. The term "LSP stitching" has potential to be ambiguous. In order to reduce the ambiguity, both meanings of the usage of the term "LSP stitching" are clarified below. One meaning of the term "LSP stitching" is as defined in [GMPLS-STITCHING].

An alternate use of "LSP stitching" as occurs for inter-AS scenarios described in [INTER-AS-VPNS]. When section 10 in [INTER-AS-VPNS]

refers to either of the two following statements it is essentially describing "LSP stitching" in the data plane:

- In "b)": "This procedure requires that there be a label switched path leading from a packet's ingress PE to its egress PE."
- In "c)": "Like the previous procedure, it requires that there be a label switched path leading from a packet's ingress PE to its egress PE."

Labeled data traffic flowing over e2e MPLS LSPs, that have been setup by stitching together segments, would benefit from having the entropy label be included in it. The specification of [EL-RFC] can be optimized for usage in such environments.

This document specifies optimizations for "LSP stitching" which are applicable to the latter meaning of the term "LSP stitching". For terminology sake, this document shall refer to that latter case as "LSP concatenation".

LSP hierarchy is defined in [MPLS-ARCH] and [GMPLS-HIER]. Usage of entropy label in LSP hierarchies has some peculiar practical issues that will benefit from some additional flexibility in inserting ELs for a specific layer in an LSP hierarchy.

2. Terminology

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

The following acronyms/terms are used:

e2e: End to end LSP that has been setup by concatenating together LSP segments

ECMP: Equal Cost Multi-Path

EL: Entropy Label

ELC: Entropy Label Capability or Entropy Label Capable

ELI: Entropy Label Indicator

Intrinsic ELC: Entropy label capability/capable as in [EL-RFC]. In this document, an LSP is considered to be

"intrinsically" EL-capable when the:

- ingress of the LSP has the ability to compute and PUSH the EL before PUSHing the ELI before PUSHing the LSP label; and
- egress/PHR of the LSP-segment has the ability to POP the (ELI+EL) at the egress/PHR while POPping-transport-label/ELI-is-top-label respectively.

LAG: Link Aggregation Group

LER: Label Edge Router

LSP: Label Switched Path

LSR: Label Switching Router

Notional ingress: Ingress LER for an LSP segment that is inserting the (ELI+EL) on data traffic going over an e2e LSP

Notional egress: egress LER for an LSP segment that is removing the (ELI+EL) from data traffic going over an e2e LSP

Notional LSP segment: the portion of the e2e LSP between a notional ingress and a notional egress

PHP: Penultimate Hop Popping

PHR: Penultimate Hop Router

UHP: Ultimate Hop Popping

NOTE: this document references the (ELI+EL) pair simply as EL when the presence of the ELI is of no significance for the issue being described. The presence of ELI is mandatory as per [EL-RFC] when EL is in use.

3. Key attributes of the entropy label solution: Summary from [EL-RFC]

- Transport-label-PUSHing router inserts (ELI+EL)
The (ELI+EL) insertion is done, if at all, by a router that is PUSHing the transport LSP's label.
- Ingress-LER (transport-label-PUSHing-router) inserts (ELI+EL) only if the PHR/egress has signaled ability to strip it off.
- Transport-label-POPping router POPs (ELI+EL) PHR/egress of the

LSP is responsible for POPing off the (ELI+EL) after it has been exposed as the top label on the packet due to POPing the transport label. The removal of the (ELI+EL) is done either when the ELI is the top label; or when the ELI is next label below the top label being POPed.

- Max-payload size for the LSP gets reduced by 8 bytes after the insertion of the (ELI+EL).

4. Problems and Motivation

[EL-RFC] defines EL signaling/usage suitable for single-segment LSPs.

[EL-RFC] does not explicitly specify the EL-signaling-interaction between concatenated LSP segments. Similarly, peculiarities in the data-plane related to LSP concatenation need further specification. Likewise, the signaling and data-plane peculiarities for using EL over LSP hierarchies could be further specified.

It is desirable to get the benefits of EL even for concatenated LSPs.

Certain aspects peculiar to concatenated LSPs need additional handling to increase the applicability of [EL-RFC]. [EL-RFC] needs to be extended to define the behavior for LSP concatenation and LSP hierarchies (tunneling) when using EL.

The sub-sections below list the specific additional requirements for making entropy label more deployable when used with LSP concatenation, and LSP hierarchy.

4.1 EL applicability for seamless MPLS

The seamless MPLS architecture relies on the use of LSP concatenation and hierarchy to signal an e2e LSP between access-nodes, such that the e2e LSP is going over aggregation/transport/cores nodes.

The signaling of such e2e LSPs is enabled by using the concatenation/hierarchy mechanisms that exist, using [LU-BGP]/LDP/RSVP.

The rules of section 5 provide a general-purpose way for the use of ELs across e2e LSPs by defining:

- the rules of ELC propagation at concatenation points;
- the data-plane guidelines for the concatenation point LSR; and

- the data-plane guidelines for LSP hierarchies for inserting (ELI+EL) at ingress LER of an LSP in an LSP hierarchy.

4.2 EL for LSP concatenation

A concatenated e2e LSP might be stitched from greater than 2 segment LSPs (along the length of the e2e LSP), with 2 LSPs forming the stitch at each concatenation point.

An LSP segment is considered to be intrinsically EL capable when it supports the attributes summarized in section 3.

Some of the segment LSPs in the e2e LSP may intrinsically support EL and some may not. So, the e2e LSP may not intrinsically support EL from end to end. However, to obtain the benefits of EL for concatenated LSPs, it is important that an EL should be present on the data packets traversing as many segments of the e2e LSP as is possible within data plane abilities of the routers on the way.

In using EL with LSP concatenation, the aims are BOTH of the following:

- a. Get EL benefits wherever possible: on all segments where possible. Just because a given segment does not support EL is not a reason to deny EL benefits to other segments of the e2e LSP.
- b. Not run into data-plane problems where an intermediate-segment whose ingress LER can not look deeper to remove EL when the subsequent segment does not support EL.

- Independent setup of LSP segments:

LSP concatenation typically relies on LSP segments that have been independently setup. In an e2e LSP (made of concatenated segments), it is unlikely that all of the concatenation points (i.e., segment LSP end points) as well as the ultimate ingress and ultimate egress support EL as defined in section 3.

However, there would be individual LSP segments that would completely satisfy the requirements of section 2 (i.e. are intrinsically EL capable). This document describes how EL may be used for (portions of) the e2e LSP while still working within the framework for [EL-RFC].

S---A---B---C---D

In the above topology, for an e2e LSP from S to D, the segments AB

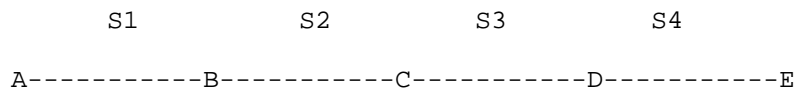
and CD could be intrinsically EL capable while the segments SA, & BC may not be. For data traffic going over the LSP from S to D, using EL on the segments AB and CD would be beneficial for load-balancing over LAGs/ECMP.

- Dealing with different protocols being used to setup the segments of the e2e LSP.

4.2.1 Spectrum of EL usage behaviors required to be supported for concatenated LSPs

To cater for an incremental deployment of intrinsically-ELC routers in a network, the multiple different modes for EL use with LSP concatenation need to be supported.

The spectrum of supported behaviors are listed below by referencing the following diagram.



LSP segments S1, S2, S3, S4 are between LERs A/B/C/D/E. There may or may not be other routers between the per-segment ingress<->egress LERs.

Transport LSP signaling protocol: could be any: LDP/RSVP/([LU-BGP] tunneled over RSVP/LDP).

4.2.1.1 Entropy label for per-segment LSP

Each of the segments will have their independent EL capability based on BOTH the:

- Per-segment ingress having the ability to insert the EL.
- Per-segment egress (or PH router) having the ability to strip the EL.

This is very similar to [EL-RFC] with the additional data-plane rule of section 5.2.2.1 "A. Rationalizing EL for the outgoing LSP segment:".

Reasoning for why per-segment EL may be attractive for certain use scenarios:

Opportunity to get benefits on those segments where EL benefits are available. Even though the e2e LSP may not support ELC, this allows

the EL benefits on those segments that are EL-capable.

4.2.1.2 Entropy label for notional-segment-LSP(s)

In the case of concatenated LSPs, it is useful to:

- Insert EL at first per-segment ingress LER (per-segment ingress LER closest to the e2e ingress LER) that has the ability to insert EL.
- Carry the EL on the data packets as far along the concatenated LSP

as the last per-segment egress LER that ability to strip the EL on a series of contiguous EL-supporting segments.

The above is enabled by the rules of section "5.2.1.1 Signaling ELC at concatenation points (Translation rules)".

The benefit of using EL with notional-segment LSPs:

An operator might be able to use EL for the MPLS traffic on its path to a concatenation point even though the concatenation-point router (or its PHR) itself may not have the data-plane capabilities required as in [EL-RFC].

Additionally, even if the concatenation-point (or its PHR) do have the data-plane capabilities of [EL-RFC], it is just more efficient to forward the data packets without having to strip the EL and then reinsert the EL when the downstream segment is also intrinsically ELC.

4.2.1.3 Entropy label for e2e LSP

This correspond to having the notional-LSP and the e2e LSP being the same.

This is covered by the rules of section 5.2.1.1 "Signaling ELC at concatenation points (Translation rules):" with the additional requirement that the data-plane be exactly the same as [EL-RFC], i.e.

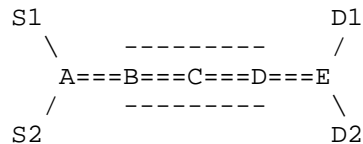
- the (ELI+EL) insertion is done by a label PUSHing router,
- the (ELI+EL) POP is done by the PHR/egress for the e2e LSP.

4.3 EL for LSP hierarchy

For the purpose of highlighting the problem to be addressed and the resultant requirements to be met, the following diagram is presented

as an example.

Let there be an LSP hierarchy with the ingress for the different levels of LSP hierarchy being at different routers, such that each LSP in the hierarchy is intrinsically EL capable. The individual LSPs in the hierarchy could be a single-segment LSP or a concatenated e2e LSP.



In the above topology, let there be the following LSPs:

- L1: B->D
- L2: A->E, tunneled through LSP L1
- L3: S1->D1, tunneled through LSP L2
- L4: S2->D2, tunneled through LSP L2

All of the LSPs above are assumed to be intrinsically EL capable.

4.3.1 Possibility of unnecessary reduction of max-payload of the LSP:

Even though the aim of using EL is to get better load-balancing support, in some cases the insertion of (ELI+EL) may unnecessarily reduce the effective payload of an LSP.

In above diagram, as per [EL-RFC] for a data packet on LSP L3, the insertion of (ELI+EL) for each of the 3 LSPs: L1, L2 and L3 is not explicitly prohibited. As a result it is possible that the packet on LSP L3, might end up with 3 (ELI+EL)s (one for each LSP level in the hierarchy) thus reducing the effective payload of the LSP L3. Likewise for L4. The presence of the (ELI+EL) for the outer LSPs L1 and L2 is not strictly useful for load-balancing the data traffic on the LSPs L3 and L4.

The solution for this issue is presented in section 5.3.2: it relies on inserting the (ELI+EL) in the context of only 1 LSP in a hierarchy.

This issues results in the following requirement for EL usage in the presence of LSP hierarchies:

- Desirability of having a single (ELI+EL) on data packets over an LSP hierarchy: The LSP for which the (ELI+EL) is inserted, is preferably the innermost intrinsically EL-capable LSP, as the

notion of a user-flow is more specifically indicated by fields deeper inside the packet headers. Having an EL be present deeper in the packet provides load-balancing benefits of EL for the traversal of the packet across a longer stretch of the network.

If there is to be only 1 (ELI+EL) in the label stack, it imposes an additional data plane requirement on the ingress LER as described in section 5.3.2.

4.3.2 Possibility of EL being non-usable for load-balancing: Even though the aim of using EL is to get better load-balancing, in some cases the insertion of (ELI+EL) may actually offer no load-balancing benefits at all. Whether the presence of an EL offers load-balancing benefits on a given transit router, depends on:

- whether the transit router has a LAG or an ECMP as an outgoing interface for the LSP, AND
- whether the forwarding ASICs of the transit routers have the ability to include the EL (appearing at a specific position in the label stack) in the hash computation, either by:
 - + looking up the ELI and then picking the EL, or
 - + computing the hash on the maximum number of labels that it can pick from the label-stack for hash-computation which happens to also include the EL.

When the EL on a packet is outside the portion-of-the-label-stack that the ASIC of a transit router can use for hash computation, the forwarding hardware may include only the top few labels or the bottom few labels in the hash computation. This may prevent the inclusion of EL for hash-computation.

In the above diagram, for a data packet going over LSP L3 let the issue of section 4.3.1 have been resolved by the router S1 inserting the (ELI+EL) underneath the label for LSP L3 and none of the other routers inserting the (ELI+EL). When this data packet arrives at router C, its label stack looks thus:

Label-LSP1		Label-LSP2		Label-LSP3		ELI		EL
Top-label				->				Bottom label

Let's say that the router C is able to include only the top 4 labels in a label stack for the hash-computation due to the ability of its forwarding ASICs.

So, the router C is not able to get the benefit of the presence of the EL in the data packet. If the only ECMP/LAG in this network is the link between C&D, then the presence of the EL serves no purpose

for the above network example and it ends up reducing the payload capacity of the LSPs L3 and L4 by 8 bytes.

This example could be further generalized in the case of seamless MPLS, where there may be deeper LSP hierarchies.

A transit router that has the ability to hash on an EL (based on its depth in the label stack) does not have multiple paths; while another router that has multiple paths and the ability hash on the EL (appearing at a specific depth in the label stack) is unable to do so as the EL appears outside the depth of the label stack that may be included in the hash. In neither of the foregoing cases is the presence of an EL helpful.

This translates into a requirement for EL: Flexibility in choice of LSP tunnel for which EL is inserted:

There is a need to have a way by which to include an EL underneath a specific label in a label-hierarchy based on it serving the most useful purpose (i.e. taking into consideration location of multiple-forwarding-paths and stack-depth-concerns).

[EL-RFC] has no way of influencing the insertion of (ELI+EL) at a certain LSP level in the stack. Thus, there is a need for a mechanism by which one of the many intrinsically-EL-capable LSPs in an LSP hierarchy could be picked for inserting the (ELI+EL).

5. EL for LSP concatenation/hierarchy

5.1 Additional EL abstractions: specific to LSP concatenation/hierarchy

Given the previous sections, following additional abstractions need to be defined to make EL more useful for LSP concatenation and hierarchy.

5.2 New abstractions: EL applicability for LSP concatenation

5.2.1 Signaling

New abstractions need to be defined to handle the differences in the use of ELs for concatenated-LSPs as compared to their use for single-segment LSPs.

The differences are:

- Notion of ingress for EL insertion:
(ELI+EL) insertion might not necessarily be done by a label-PUSHing router. A concatenation point where the label is being swapped might do the (ELI+EL) insertion, and serves as a

"notional ingress".

- Notion of egress for EL:

"Notional-egress" might not be the segment egress for the segment of the notional-ingress.

Even though certain concatenation-points (segment LERs) might not support POPing (ELI+EL), it may be acceptable to let the (ELI+EL) continue to be on the packet since the egress of a subsequent segment has the capability to POP the (ELI+EL) (which may not necessarily be along with POPing the transport label). A "notional-ingress and notional-egress" pair might actually be the segment-ingress and segment-egress for different LSP segments that are part of the same e2e LSP.

The portion of the concatenated e2e LSP, between a notional-ingress and a notional-egress is referred to as the "notional-LSP-segment" in this document.

As a packet traverses an e2e LSP, it may have an (ELI+EL) imposed on it and then removed at different routers.

It is desirable for there to not be more than one instance of an (ELI+EL) to appear on a packet at any given time. However, the insertion followed by removal of an (ELI+EL) may happen more than once as the packet traverses the e2e LSP. Each router doing the (ELI+EL) insertion is the notional-ingress and each router doing the (ELI+EL) removal is the notional-egress (or notion EL-stripping-PH-router).

Thus, there may be more than 1 "notional ingress" for EL insertion, and there may be more than 1 "notional egress" for EL removal.

For each notional "ingress ingress", there will be a "notional egress" with the "notional ingress"es and "notional egress"es alternating along the path of the e2e LSP when there are more than 1 notional ingress and egress for an e2e LSP.

In the simplest case, this boils down to the case of there being just one notional ingress and one notional egress; and the notional ingress coincides with the e2e ingress, and the notional-egress coincides with the e2e egress. That is the case that [EL-RFC] addresses.

Separation of control/data-plane implies that certain routers

- Might be running software that supports signaling ELC and understanding an egress' ELC.
- However, might not have the capability to insert (ELI+EL).

Such routers should not be allowed to play a spoil-sport in preventing EL benefits for traffic traversing over them via concatenated LSPs. In other words, such routers can not act as notional-ingress or notional-egress. However, the presence of such per-segment ingress/egress routers on the path of a notional segment-LSP should not prevent the notional segment-LSP from benefiting from the use of EL.

5.2.1.1 Signaling ELC at concatenation points (Translation rules)

The rules for propagating ELC, at concatenation points, from a downstream segment LSP to an upstream segment LSP are listed in this section.

There is benefit in propagating ELC across concatenation points is to not have to re-compute the EL at different segment ingress for those segments that are EL capable, including when the LSP segments have been setup using different protocols.

Additionally, in certain cases it should be possible to get benefits of (ELI+EL) on LSP segments that are not "intrinsically EL capable", where the lack of "intrinsic EL capability" is due to:

- The per-segment ingress does not support EL insertion.
- The per-segment PHR/egress does not support EL POPing.

However, such a concatenation point might support ELC signaling.

At a concatenation point, when 2 LSP segments: L1 (incoming LSP) and L2 (the outgoing LSP) are being concatenated, the following rules should be followed by concatenation point in signaling its ELC.

A. Segment-egress:

1. A segment-egress signals ELC for an LSP-segment L1 when:
 - a. The segment-egress is intrinsically ELC, or
 - b. When it is not intrinsically-ELC, however segment-egress for LSP-segment L2 (downstream of L1)- for which this concatenation-point is segment-ingress - is signaling ELC.
[This handles the case: Support the signaling even though it may not support the data-plane behavior.]
2. A segment-egress MUST NOT signal ELC if BOTH of the following are true:
 - a. It is also segment-ingress for another LSP-segment whose segment-egress is not signaling ELC.
 - b. This router does not have the ability to remove an (ELI+EL) inserted by the segment-ingress for the LSP-segment for which this router is the segment-egress.

B. Segment-ingress:

The following is relevant only for RSVP as defined in [EL-RFC]. When this router acting as segment-egress for an LSP L1 (that is concatenated to downstream LSP L2) is signaling ELC for L1, then this router must signal ELC in its Path messages using the mechanism defined in [EL-RFC].

This is relevant only in the context of bidirectional LSPs.

5.2.2 Data plane aspects

5.2.2.1 LSP concatenation: Differing EL dispositions

At a concatenation point, when 2 LSP segments: L1 (incoming LSP) and L2 (the outgoing LSP) are being concatenated, the following rules should be followed by the concatenation point in its data-plane behavior.

A. Rationalizing EL for the outgoing LSP segment:

When the LSP segments L1 and L2 differ in their ELC, the concatenation point router needs to take the following data-plane actions depending on its role for the e2e LSP:

a. Notional egress behavior:

When L1 intrinsically supports ELC and L2 does not, then the concatenation-point router must remove the (ELI+EL), if present under top label, from the received data packets before effectively SWAPing the top label. In other words, in the presence of the ELI, the operations performed should be:

```
POP(IncomingLabel), POP(ELI+EL), PUSH(OutgoingLabel)
    or alternately:
POP, POP, SWAP(OutgoingLabel)
```

Translation rule "A 2" of section 5.2.1.1 would have ensured that the above is doable at the concatenation point.

b. Notional ingress behavior:

When L1 does not intrinsically support ELC and L2 does, then the concatenation point router must POP the incoming label, insert (ELI+EL) before PUSHing the label for the LSP segment L2.

The label operations performed would be:

```
POP(IncomingLabel), PUSH(EL), PUSH(ELI), PUSH(OutgoingLabel),
    or
SWAP(EL), PUSH(ELI), PUSH(OutgoingLabel)
```

c. Implicit notional ingress behavior:

When L1 is intrinsically ELC and so is L2, the arriving data traffic should already have (ELI+EL) on it.

However, it is possible that due to local configuration on the notional-ingress, (ELI+EL) is not being inserted. In that case, traffic arriving on L1 will not have (ELI+EL) on it.

In that case, this concatenation-point is the "implicit notional ingress" and it should insert (ELI+EL) just as if it were a "notional ingress".

B. Preventing multiple (ELI+EL) pairs underneath a given forwarding label in the stack:

A segment-ingress that is intrinsically-EL-capable should have the ability to inspect received data packets to check whether an (ELI+EL) already exists on the data packet underneath the top label.

Not causing multiple ELs on a data packet:

When both the LSP segments L1 and L2 support ELC, the concatenation point router SHOULD insert an (ELI+EL) only if the incoming packet does not contain an (ELI+EL) underneath the top label. In that case, the label operations are as below:

POP(IncomingLabel), PUSH(ELI+EL), PUSH(OutgoingLabel)

If the incoming packet already contains an (ELI+EL) underneath the top label, an additional (ELI+EL) MUST NOT be inserted on the packet underneath the top label that is being effectively SWAPed.

This prevents a segment ingress from inserting an (ELI+EL) when the notional ingress has already inserted an (ELI+EL).

C. Rationalizing EL insertion (at concatenation-point) for LSP hierarchy:

A concatenation point router that is intrinsically-EL-capable should have the ability to inspect received data packets to check whether an (ELI+EL) already exists, underneath any label in the label-stack.

If the router has such a ability, then this router MUST NOT insert an (ELI+EL) as in "A a" above.

This helps to prevent multiple (ELI+EL)s on the packet inserted (at a concatenation point) in the context of different transport labels in a label hierarchy.

D. Notional ingress role change at a router:

This role can change due to local configuration on the router or due to segment egress starting/stopping to signal ELC possibly due to a configuration change at the segment egress or due to a configuration change at this router.

When this router becomes a notional ingress, it reacts to the change as in "A b" above.

When this router stops being a notional ingress, this router stops inserting the (ELI+EL) underneath the top label that this router is

 SWAPing (if this router is concatenation point), or
 PUSHING (if this router is e2e ingress).

E. Notional egress role change at a router:

This role can change due to local configuration on the router or due the egress of a downstream concatenated LSP segment starting to signal ELC.

When this router becomes a notional egress, it reacts to the change as in "A a" above.

When this router stops being a notional egress, this router stops performing the label operation described in "A a" above. Instead this router now starts to simply SWAP the top label.

5.3 New abstractions: EL applicability for LSP hierarchy

5.3.1 Management plane:

Moving the (ELI+EL) underneath a different LSP's transport label:

There are 2 ways to handle the issue of section 4.3.2:

- Configuration at the ingress LER: a configuration option should exist by which an operator can disable the insertion of (ELI+EL) on a per-LSP basis. The specific level in the LSP hierarchy for which to enable this configuration is based on operator knowledge based on:
 - * Knowledge of transit routers that need EL benefits : those routers that have a multi-path (LAG or LSP ECMP) as egress interface.
 - * The label hashing abilities of such routers: information

about the specific number of labels in the label-stack that can be used in the hash computation; and any constraints about the position of the labels that can be used for computation when the label stack is larger than a certain ASIC-specific threshold.

- Configuration-based rewrite of the label stack at the ingress LER of an intrinsically-EL-capable LSP:

An operator will know the forwarding characteristics (with regards to the number of labels that can be included in the hash computation) of the transit routers across the path of the e2e LSP that is part of an LSP hierarchy.

By making such a configuration, the operator can ensure that the EL will appear in the label stack such that all transit routers shall be able to include the as part of the hash computation.

The configuration would cause the label stack of the outgoing packet to have its extant (ELI+EL) removed, and an (ELI+EL) inserted just underneath the label of the LSP for which this ingress LER is setup to insert (ELI+EL).

5.3.2 Data plane aspects

Preventing insertion of multiple (ELI+EL)s:

At an ingress LER, the router SHOULD not insert an (ELI+EL) for an LSP if the packet already contains an ELI.

This ensures that for a data packet on a hierarchy of LSPs, there will be only 1 instance of an (ELI+EL). This helps to prevent the issue of section 4.3.1.

This also ensures that when multiple LSPs in an LSP hierarchy are intrinsically-EL-capable, the (ELI+EL) will be inserted just underneath the transport label of the innermost LSP in the hierarchy. However, based on section 5.3.1 there is a way by which to modify the level in the LSP hierarchy for which an (ELI+EL) is inserted.

A more specific case of this is already covered in section "5.2.2.1 C. Rationalizing EL insertion (at concatenation-point) for LSP hierarchy:".

6. Use-cases

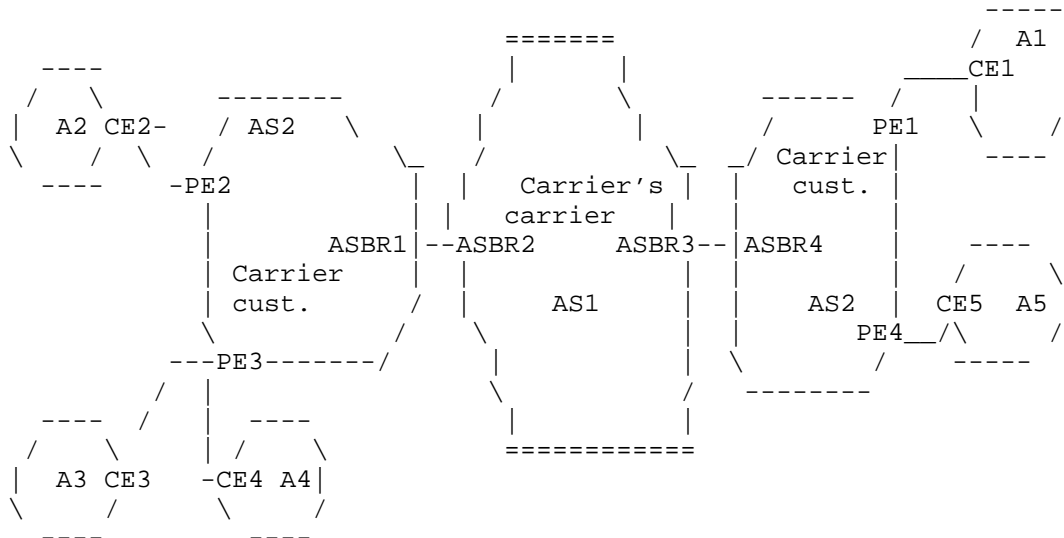
In this document, the definition of LSP-concatenation refers to those cases where a label advertised by one label distribution protocol being:

- removed at one router (by PH POP) followed by a label PUSH for a label distributed by another protocol at the router downstream of the PH router of the previous protocol's LSP. Such a router which is PUSHing a label for a subsequent protocol is referred to as a concatenation-point router in this document.
- SWAPed at a router for a label distributed by another protocol is also referred to as a concatenation-point in this document.

The list of use-cases of this draft stems from the following possible optimizations:

- A. Not having to insert/remove (ELI+EL) multiple times along an e2e labeled path, due to the EL capability not getting signaled e2e. In other words, not having to remove (ELI+EL) at a concatenation point only to re-insert it.
The lack of e2e EL capability signaling could be either due to lack of support; or due to a label advertised by one label distribution protocol being removed at one router (which also causes the removal of (ELI+EL)) and a label distributed by a subsequent router being PUSHed along with (ELI+EL) at that router.
- B. On an e2e labeled transport-LSP path, it may be possible to get the load-balancing benefits of EL on (segment of) the e2e LSP even though not every concatenation point router (as defined above) may intrinsically support EL for the LSP terminating at it.

6.1 Carrier of carrier L3VPN



A<n> = Customer site n
 CE = Customer Edge Device
 PE = Provider Edge Router

In the above figure, the "carrier's carrier" is providing L3VPN service to a carrier customer (carrier cust.) is itself an L3VPN provider.

Let the sites A<n> be the sites of the same L3VPN.

In order to provide L3VPN service to the sites A<n>, there is effectively an e2e LSP between each pair of PEs. For PEs in the same carrier customer site, the e2e LSP is an RSVP or LDP LSP. eg. Between PE2 and PE3.

For PEs that are across the carrier-customer's core, there is an e2e LSP created by advertising a BGP label for the remote PE's loopback address. The BGP label advertised from ASBR2 to ASBR1 rides-on top of the RSVP or LDP label in the carrier's-carrier core.

eg. For having an e2e LSP from PE1 to PE2, a BGP label is advertised for PE1's loopback into the carrier customer's site on the left. This label could be dealt with by ASBR1 in two ways:

- Advertising it into LDP in the carrier customer's site (on the left), or
- By advertising it over an iBGP session to PE2.

In the former case (LDP advertising a FEC for PE1), this document makes possible for ASBR1 to not have to remove the EL (inserted by PE2) and let it be removed by either a concatenation point (ASBR2 or ASBR3 or ASBR4) or the egress PE1. This is facilitated by the

translation rules of section 5.2.1.1. The same also facilitates traffic with EL to be carried over concatenation points such that the EL is eventually removed by the last-EL-capable concatenation point or the EL capable e2e egress.

Each carrier (carrier's carrier; and carrier-customer) will have LAGs and LDP ECMP paths in its network.

6.2 Inter-AS L3VPN: Option C

Option C is conceptually similar to CoC L3VPN from a point of view of setting up the e2e LSP. Therefore, similar EL use-cases also exist for Option C.

This applies for both L3VPN and also BGP-VPLS.

7. Manageability

There are no new MPLS OAM issues opened up by this specification. Any MPLS manageability are the same as those inherited from [EL-RFC] and addressing those is outside the scope of this document.

8. Security considerations

Security considerations as listed in section 9 of [EL-RFC] apply.

9. Acknowledgments

Many thanks to Adrian Farrel for his inputs on the stitching scenarios, and suggesting editorial improvements.

Thanks to the EL team (Sudharsana Venkataraman, Nitin Singh, Ramji Vijayaraghavan, Jie Yan, Abhishek Tripathi) for discussions on some of these topics.

10. IANA considerations

None.

11. References

11.1 Normative References

- [EL-RFC] Kompella, K., Drake, J., Amante, S., Henderickx, W., L. Yong, "The Use of Entropy Labels in MPLS Forwarding", RFC-6790, November 2012.
- [GMPLS-HIER] Kompella, K., Y. Rekhter, "Label Switched Paths (LSP) Hierarchy with Generalized Multi-Protocol Label Switching (GMPLS)", RFC-4206, October 2005.
- [MPLS-ARCH] Rosen, E., Viswanathan, A., R. Callon, "Multiprotocol Label Switching Architecture", RFC-3031, January 2001.
- [S-MPLS] Leymann, N., Decraene, B., Filsfils, C., Konstantynowicz, M., Steinberg, D., "Seamless MPLS Architecture", draft-ietf-mpls-seamless-mpls-07, October 2012. (work-in-progress)
- [INTER-AS-VPNS] Rosen, E., Y. Rekhter, "BGP/MPLS IP Virtual Private Networks (VPNs)", RFC4364, February 2006.
- [GMPLS-STITCHING] Ayyangar, A., Kompella, K., Vasseur, JP., A. Farrel, "Label Switched Path Stitching with Generalized Multiprotocol Label Switching Traffic Engineering (GMPLS TE)", RFC 5150, February 2008.

11.2 Informative References

- [ISSUE-DEEP] K. Kompella, "Deep Label Stacks", <http://tools.ietf.org/agenda/84/slides/slides-84-mpls-15.pdf>, August 2012
- [LU-BGP] Rekhter, Y., E. Rosen, "Carrying Label Information in BGP-4", RFC-3107, May 2001.

Authors' Addresses

Ravi Singh
Juniper Networks
1194 N. Mathilda Ave.
Sunnyvale, CA 94089
US

EMail: ravis@juniper.net

Yimin Shen
Juniper Networks
10 Technology Park Drive
Westford, MA 01886
US

EMail: yshen@juniper.net

John Drake
Juniper Networks
1194 N. Mathilda Ave.
Sunnyvale, CA 94089
US

EMail: jdrake@juniper.net

MPLS Working Group
Internet-Draft
Intended status: Standards Track
Expires: January 4, 2015

Kamran Raza
Nobo Akiya
Carlos Pignataro
Cisco Systems, Inc.
July 3, 2014

IPv6 Router Alert Option for MPLS OAM
draft-raza-mpls-oam-ipv6-rao-00

Abstract

RFC4379 defines the MPLS LSP Ping/Traceroute mechanism, in which the Router Alert option must be set in the IP header of the MPLS Echo Request messages, and may conditionally be set in the IP header of the MPLS Echo Reply messages. While a generic "Router shall examine packet" Option Value is used for the IPv4 Router Alert Option (RAO), there is no generic Router Alert Option Value defined for IPv6 that can be used. This document allocates a new generic IPv6 Router Alert Option Value that can be used by MPLS OAM tools, including the MPLS Echo Request and MPLS Echo Reply messages for MPLS IPv6.

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 4, 2015.

Copyright Notice

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

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents

carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction	2
2. Specification of Requirements	3
3. IPv6 Router Alert Option (RAO) Value for MPLS OAM	3
4. IANA Considerations	3
5. Security Considerations	3
6. Acknowledgments	3
7. References	3
7.1. Normative References	4
7.2. Informative References	4
Authors' Addresses	4

1. Introduction

A commonly deployed MPLS OAM tool is LSP Ping/Traceroute [RFC4379] which is used to diagnose MPLS networks. The LSP Ping/Traceroute specification [RFC4379] requires the use of Router Alert option in the IP header. For example, the section 4.3 of [RFC4379] states that IP Router Alert option MUST be set in the IP header of an MPLS Echo Request message. Similarly, the section 4.5 states that IP Router Alert option MUST be set in the IP header of an MPLS Echo Reply message if the Reply Mode in the echo request is set to "Reply via an IPv4/IPv6 UDP packet with Router Alert".

[RFC2113] defines a generic Option Value 0x0 for IPv4 Router Alert Option (RAO) that is used by LSP Ping and LSP Traceroute for MPLS IPv4. However, currently there is no generic IPV6 Router Alert code point defined that can be used by LSP Ping and LSP Traceroute for MPLS IPv6. Specifically, [RFC2711] defined the router alert for a general IPv6 purpose but required the Value field in the router alert option to indicate a specific reason for using the router alert option. Because there is no defined value for MPLS LSP Ping/Traceroute use or for general use, it is not possible for MPLS OAM tools to use the IPv6 Router Alert mechanism.

As vendors are starting to implement MPLS on IPv6 control plane (e.g., [I-D.ietf-mpls-ldp-ipv6]), there is a need to define and allocate such a code point for IPv6 in order to comply with [RFC4379]. This document defines a new IPv6 Router Alert Option Value that can be used by MPLS OAM tools, including the MPLS Echo Request and MPLS Echo Reply messages for MPLS IPv6.

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 [RFC2119].

3. IPv6 Router Alert Option (RAO) Value for MPLS OAM

This document defines a new option value (TBD1) for the IPv6 Router Alert Option (RAO) to alert transit routers to examine the packet more closely for MPLS OAM purposes. This code point is used by any MPLS OAM application that requires their packets to be examined by a transit router.

In the scope of this document, this code point will be used by the MPLS Echo Request and MPLS Echo Reply for its IPv6 messages as required by [RFC4379].

4. IANA Considerations

This document defines a new code point (value TBD1) for IPv6 Router Alert option to alert transit routers to examine the packet the MPLS OAM purpose. IANA is requested to assign a new code point under its "IPv6 Router Alert Option Values" registry defined by [RFC5350] and maintained in [IANA-IPv6-RAO]. The new code point is as follows:

value	Description	Reference
-----	-----	-----
TBD1	MPLS OAM	[document.this]

5. Security Considerations

This document introduces no new security concerns in addition to what have already been captured in [RFC4379] and [RFC6398].

6. Acknowledgments

The authors would like to thank George Swallow, Ole Troan, Bob Hinden, Faisal Iqbal, and Mathew Janelle for their useful input.

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.
- [RFC2711] Partridge, C. and A. Jackson, "IPv6 Router Alert Option", RFC 2711, October 1999.
- [RFC4379] Kompella, K. and G. Swallow, "Detecting Multi-Protocol Label Switched (MPLS) Data Plane Failures", RFC 4379, February 2006.
- [RFC5350] Manner, J. and A. McDonald, "IANA Considerations for the IPv4 and IPv6 Router Alert Options", RFC 5350, September 2008.
- [RFC6398] Le Faucheur, F., "IP Router Alert Considerations and Usage", BCP 168, RFC 6398, October 2011.

7.2. Informative References

- [I-D.ietf-mpls-ldp-ipv6] Asati, R., Manral, V., Papneja, R., and C. Pignataro, "Updates to LDP for IPv6", draft-ietf-mpls-ldp-ipv6-12 (work in progress), February 2014.
- [IANA-IPv6-RAO] IANA, "IPv6 Router Alert Option Values", <<http://www.iana.org/assignments/ipv6-routeralert-values/ipv6-routeralert-values.xhtml>>.
- [RFC2113] Katz, D., "IP Router Alert Option", RFC 2113, February 1997.

Authors' Addresses

Kamran Raza
Cisco Systems, Inc.
2000 Innovation Drive
Kanata, ON K2K-3E8
Canada.
Email: skraza@cisco.com

Nobo Akiya
Cisco Systems, Inc.
2000 Innovation Drive
Kanata, ON K2K-3E8
Canada.
Email: nobo@cisco.com

Carlos Pignataro
Cisco Systems, Inc.
7200-12 Kit Creek Road
Research Triangle Park, NC 27709
USA.
Email: cpignata@cisco.com

MPLS Working Group
Internet-Draft
Intended status: Standards Track
Expires: October 24, 2014

Tarek Saad, Ed.
Rakesh Gandhi, Ed.
Zafar Ali
Cisco Systems, Inc.
Robert H. Venator
Defense Information Systems Agency
Yuji Kamite
NTT Communications Corporation
April 22, 2014

Reoptimization of Point-to-Multipoint Traffic Engineering
Loosely Routed LSPs
draft-tsaad-mpls-p2mp-loose-path-reopt-02

Abstract

This document defines Resource Reservation Protocol - Traffic Engineering (RSVP-TE) signaling extensions for reoptimizing loosely routed Point-to-Multipoint (P2MP) Traffic Engineered (TE) Label Switched Paths (LSPs) in an Multi-Protocol Label Switching (MPLS) and/or 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."

Copyright Notice

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

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

Table of Contents

1. Introduction	3
2. Terminology	4
2.1. Abbreviations	4
2.2. Nomenclatures	5
2.3. Conventions used in this document	5
3. Signaling Procedure For Loosely Routed P2MP-TE LSP Reoptimization	5
4. RSVP Signaling Extensions	6
4.1. P2MP-TE Tree Re-evaluation Request Flag	6
4.2. Preferable P2MP-TE Tree Exists Path Error sub-code	6
5. Compatibility	7
6. Security Considerations	7
7. IANA Considerations	7
7.1. P2MP-TE Tree Re-evaluation Request Flag	8
7.2. Preferable P2MP-TE Tree Exists Path Error sub-code	8
8. Acknowledgments	8
9. References	9
9.1. Normative References	9
9.2. Informative References	9
Author's Addresses	10

1. Introduction

This document defines Resource Reservation Protocol - Traffic Engineering (RSVP-TE) [RFC2205] [RFC3209] signaling extensions for reoptimizing loosely routed Point-to-Multipoint (P2MP) Traffic Engineered (TE) Label Switched Paths (LSPs) [RFC4875] in an Multi-Protocol Label Switching (MPLS) and/or Generalized MPLS (GMPLS) networks.

A P2MP-TE LSP is comprised of one or more source-to-leaf (S2L) sub-LSPs. A loosely routed P2MP-TE S2L sub-LSP is defined as one whose path does not contain the full explicit route identifying each node along the path to the egress node at the time of its signaling by the ingress node. Such an S2L sub-LSP is signaled with no Explicit Route Object (ERO) [RFC3209], or with an ERO that contains at least one loose hop, or with an ERO that contains an abstract node that is not a simple abstract node (that is, an abstract node that identifies more than one node). This is often the case with inter-domain P2MP-TE LSPs where Path Computation Element (PCE) is not used [RFC5440].

As per [RFC4875], an ingress node may reoptimize the entire P2MP-TE LSP by resignaling all its S2L sub-LSP(s) or may reoptimize individual S2L sub-LSP(s) i.e. individual destination(s).

[RFC4736] defines RSVP signaling extensions for reoptimizing loosely routed P2P TE LSP(s) as follows.

- An egress border node MAY send a solicited or unsolicited PathErr with the Notify error code (25 as defined in [RFC3209]) with sub-code 6 to indicate "Preferable Path Exists" to the ingress node.
- An ingress node MAY solicit a PathErr that indicates "Preferable Path Exists" by sending a "Path Re-evaluation Request" to an egress border node by setting a flag (0x20) in SESSION_ATTRIBUTES object in the Path message.
- The ingress node upon receiving this PathErr either solicited or unsolicited initiates reoptimization of the LSP.

[RFC4736] does not define signaling extensions specific for reoptimizing entire P2MP-TE LSP tree. Mechanisms defined in [RFC4736] can be used for signaling the reoptimization of individual S2L sub-LSP(s). However, to use [RFC4736] mechanisms for reoptimizing an entire P2MP-TE LSP tree, an ingress node needs to send the query on all (typically 100s of) S2L sub-LSPs and an egress border node needs to notify PathErrs for all S2L sub-LSPs. Such a

procedure may lead to the following issues.

- An egress border node has to accumulate the received queries on all S2L sub-LSPs (using a wait timer) and interpret them as a reoptimization request for the P2MP-TE LSP tree. An egress border node may prematurely notify "Preferable Path Exists" for one or a sub-set of S2L sub-LSPs.

- When the ingress node gradually receives unsolicited PathErr notifications for individual S2L sub-LSPs, it may prematurely start reoptimizing a sub-set of S2L sub-LSPs. However, as mentioned in [RFC4875] Section 14.2, such reoptimization procedure may result in data duplication that can be avoided if the entire P2MP-TE LSP tree is reoptimized, especially if the ingress node eventually receives PathErr notifications for all S2L sub-LSPs of the P2MP-TE LSP tree.

- The ingress node may have to heuristically determine when to perform entire P2MP-TE LSP tree reoptimization versus per S2L sub-LSP reoptimization, for example, to delay reoptimization long enough to allow all PathErr(s) to be received. Once all PathErr(s) are received, the ingress node has to accumulate them to see if reoptimization of the entire P2MP-TE is necessary. Such procedures may produce undesired results due to timing related issues. This may be easily avoided by the RSVP signaling messages defined in this document.

This document defines RSVP-TE signaling extensions to query and notify the existence of a preferable path for reoptimizing loosely routed P2MP-TE LSP tree.

2. Terminology

2.1. Abbreviations

ABR: Area Border Router.

AS: Autonomous System.

ERO: Explicit Route Object.

TE LSP: Traffic Engineering Label Switched Path.

TE LSP ingress: head/source of the TE LSP.

TE LSP egress: tail/destination of the TE LSP.

2.2. Nomenclatures

Domain: Routing or administrative domain such as an IGP area and an autonomous system.

Interior Gateway Protocol Area (IGP Area): OSPF Area or IS-IS level.

Inter-area TE LSP: A TE LSP whose path transits across at least two different IGP areas.

Inter-AS MPLS TE LSP: A TE LSP whose path transits across at least two different Autonomous Systems (ASes) or sub-ASes (BGP confederations).

S2L sub-LSP: Source-to-leaf sub Label Switched Path.

2.3. 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]. The reader is assumed to be familiar with the terminology in [RFC4875] and [RFC4736].

3. Signaling Procedure For Loosely Routed P2MP-TE LSP Reoptimization

It might be preferable, as per [RFC4875], to reoptimize the entire P2MP-TE LSP by resignaling all its S2L sub-LSP(s) (Section 14.1, "Make-before-Break") or reoptimize individual S2L sub-LSP(s) i.e. individual destination(s) (Section 14.2 "Sub-Group-Based Re-Optimization").

It might be preferable to use the procedures defined in [RFC4736] to individually reoptimize the S2L sub-LSP(s) of a P2MP-TE LSP.

To reoptimize an entire P2MP-TE LSP tree, in order to query egress border nodes to check if a preferable P2MP-TE LSP tree exists, an ingress node sends a Path message with "P2MP-TE Tree Re-evaluation Request" defined in this document. An ingress node may select one or more S2L sub-LSP of the P2MP-TE LSP tree transiting through the egress border node to send the query to that egress border node.

An egress border node receiving the "P2MP-TE Tree Re-evaluation Request" checks for a preferable P2MP-TE LSP tree by re-evaluating loosely expanded paths for all S2L sub-LSP(s) of the P2MP-TE LSP. If a preferable P2MP-TE LSP tree is found, the egress border node immediately sends an RSVP PathErr to the ingress node with Error code 25 (Notify defined in [RFC3209] and Error sub-code defined in this

document "Preferable P2MP-TE Tree Exists". At this point, the egress border node does not propagate this bit in subsequent RSVP Path messages sent downstream for the re-evaluated P2MP-TE LSP. The sending of an RSVP PathErr Notify message "Preferable P2MP-TE Tree Exists" to the ingress node will notify the ingress node of the existence of a preferable P2MP-TE LSP tree.

If no preferable path for P2MP-TE LSP tree can be found, the recommended mode is for the egress border node to relay the request downstream by setting the "P2MP-TE Tree Re-evaluation Request" bit in the LSP_ATTRIBUTES object of RSVP Path message.

An egress border node may send "Preferable P2MP-TE Tree Exists" with PathErr code 25 to the ingress node to notify an existence of a preferred path for the P2MP-TE LSP tree with an unsolicited PathErr message. The egress border node may select one or more S2L sub-LSP(s) of the P2MP-TE LSP tree to send this PathErr message to the ingress node.

4. RSVP Signaling Extensions

4.1. P2MP-TE Tree Re-evaluation Request Flag

In order to query egress border nodes to check if a preferable P2MP-TE LSP tree exists, a new flag is defined in Attributes Flags TLV of the LSP_ATTRIBUTES object [RFC5420] as follows:

Bit Number (to be assigned by IANA): P2MP-TE Tree Re-evaluation
Request flag

The "P2MP-TE Tree Re-evaluation Request" flag is meaningful in a Path message of a P2MP-TE S2L sub-LSP and is inserted by the ingress node.

4.2. Preferable P2MP-TE Tree Exists Path Error sub-code

In order to indicate to an ingress node that a preferable P2MP-TE LSP tree is available, following new sub-code for PathErr code 25 (Notify Error) [RFC3209] is defined:

Sub-code (to be assigned by IANA): Preferable P2MP-TE Tree Exists
sub-code

When a preferable path for P2MP-TE LSP tree is found, the egress border node sends a solicited or unsolicited "Preferable P2MP-TE Tree Exists" PathErr notification to the ingress node of the P2MP-TE LSP.

5. Compatibility

The LSP_ATTRIBUTES object has been defined in [RFC5420] with class numbers in the form 1lbbbbbb, which ensures compatibility with non-supporting nodes. Per [RFC2205], nodes not supporting this extension will ignore the new flag defined in this document but forward it without modification.

6. Security Considerations

This document defines a mechanism for an egress border node to notify the ingress node of the existence of a preferable path. As per [RFC4736], in the case of a P2MP-TE LSP S2L sub-LSP spanning multiple domains, it may be desirable for a border node to modify the RSVP PathErr message defined in this document to maintain confidentiality across different domains. Furthermore, an ingress node may decide to ignore this PathErr message coming from an egress border node residing in another domain. Similarly, an egress border node may decide to ignore the path re-evaluation request originating from another ingress domain.

7. IANA Considerations

IANA maintains a name space for RSVP-TE TE parameters "Resource Reservation Protocol-Traffic Engineering (RSVP-TE) Parameters". From the registries in this name space "Attribute Flags" allocation of new flag is requested (Section 4.1).

IANA also maintains a name space for RSVP protocol parameters "Resource Reservation Protocol (RSVP) Parameters". From the sub-registry "Sub-Codes - 25 Notify Error" in registry "Error Codes and Globally-Defined Error Value Sub-Codes" allocation of a new error code is requested (Section 4.2).

7.1. P2MP-TE Tree Re-evaluation Request Flag

The following new flag is defined for the Attributes Flags TLV in the LSP_ATTRIBUTES object [RFC5420]. The numeric value is to be assigned by IANA.

- o P2MP-TE Tree Re-evaluation Request Flag:

Bit No	Attribute Flag Name	Carried in Path	Carried in Resv	Carried in RRO	Reference
TBA by IANA	P2MP-TE Tree Re-evaluation	Yes	No	No	This document

7.2. Preferable P2MP-TE Tree Exists Path Error sub-code

As defined in [RFC3209], the Error Code 25 in the ERROR_SPEC object corresponds to a Notify Error PathErr. This document adds a new sub-code as follows for this PathErr:

- o Preferable P2MP-TE Tree Exists sub-code:

Sub-code value	Sub-code Name	PathErr Code	PathErr Name	Reference
TBA by IANA	Preferable P2MP-TE Tree Exists	25	Notify error	This document

8. Acknowledgments

The authors would like to thank Loa Andersson for reviewing this document.

9. References

9.1. Normative References

- [RFC2205] Braden, R., Ed., 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.
- [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.

9.2. Informative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC4736] Vasseur, JP., Ikejiri, Y. and Zhang, R, "Reoptimization of Multiprotocol Label Switching (MPLS) Traffic Engineering (TE) Loosely Routed Label Switched Path (LSP)", RFC 4736, November 2006.
- [RFC5440] Vasseur, JP., Ed., and JL. Le Roux, Ed., "Path Computation Element (PCE) Communication Protocol (PCEP)", RFC 5440, March 2009.

Author's Addresses

Tarek Saad (editor)
Cisco Systems

Email: tsaad@cisco.com

Rakesh Gandhi (editor)
Cisco Systems

Email: rgandhi@cisco.com

Zafar Ali
Cisco Systems

Email: zali@cisco.com

Robert H. Venator
Defense Information Systems Agency

Email: robert.h.venator.civ@mail.mil

Yuji Kamite
NTT Communications Corporation

Email: y.kamite@ntt.com

Internet Engineering Task Force
Internet-Draft
Intended status: Standards Track
Expires: January 5, 2015

V. Govindan
N. Akiya
Cisco Systems
July 4, 2014

Label Switched Path (LSP) Ping
Extended Bidirectional Forwarding Detection (BFD) Discriminator TLV
draft-vgovindan-mpls-extended-bfd-disc-tlv-00

Abstract

This document defines an extended Bidirectional Forwarding Detection (BFD) discriminator TLV for the Multiprotocol Label Switching (MPLS) Label Switched Path (LSP) Ping mechanism, to allow bootstrapping of multiple BFD sessions for a given FEC.

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 5, 2015.

Copyright Notice

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

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of

publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Background	2
2. Overview	3
3. Procedures for BFD session establishment and removal using the Extended BFD TLV	3
3.1. Procedures for establishing BFD sessions	3
3.2. Procedures for removing BFD sessions	3
4. Extended BFD Discriminator TLV	4
5. Mutually Exclusive: BFD TLVs	5
6. Backwards Compatibility	5
7. Encapsulation	5
8. Security Considerations	5
9. IANA Considerations	6
9.1. Extended BFD Discriminator TLV	6
10. Acknowledgements	6
11. Contributing Authors	6
12. Normative References	6
Appendix A. Alternate format for the BFD Extended TLV	6
Authors' Addresses	8

1. Background

Bidirectional Forwarding Detection (BFD) [RFC5880] for Multiprotocol Label Switching (MPLS) Label Switched Paths (LSPs), [RFC5884], describes a mechanism to use BFD to monitor the connectivity in-band on the LSPs. The BFD session on the LSP egress is bootstrapped using the LSP Ping mechanism, defined in [RFC4379], carrying the BFD Discriminator TLV that describes the BFD discriminator of the BFD session on the LSP ingress.

The BFD Discriminator TLV and defined procedures around this TLV only allow one BFD session to be bootstrapped per <MPLS Forwarding Equivalent Class (FEC), LSP>. There are scenarios where an LSP ingress may desire to run multiple BFD sessions to monitor the connectivity on an LSP. To achieve the bootstrapping of multiple BFD sessions per FEC, a new TLV and procedures are required. Two scenarios where this is useful are described below:

- o Entropy labels help achieve load balancing of traffic belonging to the same <MPLS FEC, LSP>. It may be beneficial to track the

individual paths of the multi-path network using separate BFD sessions for each non-congruent path.

- o It may be useful to establish multiple BFD sessions for the same <MPLS FEC, LSP> to achieve BFD session redundancy, i.e. protection against false positives due to equipment or soft failures inside boxes.

2. Overview

An LSR ingress wanting to bootstrap one or more BFD sessions on an LSP is to include the Extended BFD Discriminator TLV, described in Section 4, in the MPLS echo request message for the FEC. The Extended BFD Discriminator TLV is capable of carrying multiple BFD discriminators, and each BFD discriminator is accompanied with an instance identifier. The LSR egress, upon reception of this MPLS echo request, is to create requested number of BFD sessions for the specified FEC. Each BFD session object created on the LSR ingress and the LSR egress MUST be annotated with corresponding instance identifier. BFD session procedures are to follow those described in [RFC5884].

3. Procedures for BFD session establishment and removal using the Extended BFD TLV

3.1. Procedures for establishing BFD sessions

There are at least two options possible here:

1. BFD session establishment MUST follow the procedure specified in [RFC5884].
2. The base procedure for BFD session establishment MUST be the same as that of [RFC5884]. This procedure can be enhanced by specifying additional Operation type field and Operation status field in the proposed Extended BFD Discriminator TLV. See Appendix A for a description of Operation types and Operation status codes.

3.2. Procedures for removing BFD sessions

[RFC5884] does not specify an explicit procedure for deleting BFD sessions. A few options are possible here:

1. Specify an explicit delete procedure for the BFD session using Operation types field and Operation status field through the Extended BFD TLV. See Appendix A for a description of Operation types and Operation status codes.

2. Specify a timer based deletion procedure: A new purge timer field can be introduced within the proposed Extended BFD Discriminator TLV. The ingress specifies the value for the purge timer field. Once the BFD session transitions from up to down state, the egress is to delete the session after the value specified in the purge timer field. Ed Note: This approach is an open topic for discussion.
3. No new procedure to delete a BFD session is introduced. Assumption by the egress is that BFD sessions can be deleted if corresponding FEC is deleted from the system or sometime after BFD sessions go down.

Regardless of the option chosen to proceed, all BFD sessions established with the FEC MUST be removed automatically if the FEC is removed.

4. Extended BFD Discriminator TLV

The Extended BFD Discriminator object is a new TLV that MAY be included in the MPLS echo request message. An MPLS echo request MUST NOT include more than one Extended BFD Discriminator object. The Extended BFD Discriminator object describes one or more BFD discriminators along with each having an instance identifier. An MPLS echo reply MAY include the Extended BFD Discriminator object, but MUST NOT include more than one Extended BFD Discriminator object.

Extended BFD Discriminator TLV Type is TBD1. Length is 8 or multiples of 8. Length of (8 x N) implies that there are N entries in the Value field of the Extended BFD Discriminator TLV. Each entry in the Value field of the Extended BFD Discriminator TLV has following format:

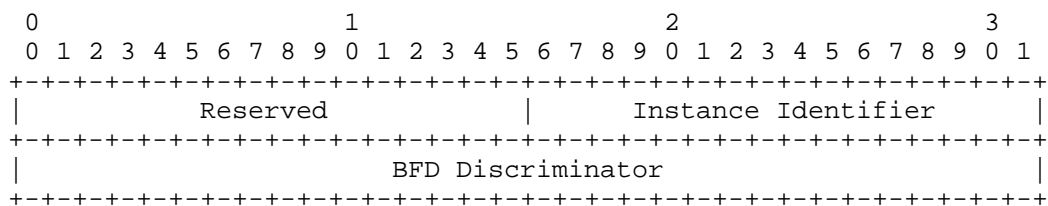


Figure 1: Extended BFD Discriminator TLV

Reserved - This field MUST be set to zero on transmit, and ignored on receipt.

Instance Identifier - An instance identifier of the BFD session. The instance identifier is a value allocated by the LSP ingress

for corresponding BFD Discriminator, and MUST be unique within the FEC on the LSP ingress node. The instance identifier MUST NOT change for the lifetime of the BFD session.

BFD Discriminator - The BFD discriminator allocated for this BFD session by the LSP ingress.

See Appendix A for a discussion on an alternate format for the TLV.

5. Mutually Exclusive: BFD TLVs

The BFD Discriminator TLV and the Extended BFD Discriminator TLV are mutually exclusive. An MPLS echo request/reply message MUST NOT include both the BFD Discriminator TLV and the Extended BFD Discriminator TLV. Reception of an MPLS echo request with both the BFD Discriminator TLV and the Extended BFD Discriminator TLV is to result in the Return Code being set to Malformed echo request received (1).

6. Backwards Compatibility

If an LSP ingress wishes to bootstrap multiple BFD sessions with the Extended BFD Discriminator TLV when an LSP already has a BFD session bootstrapped with the BFD Discriminator TLV, following procedures are RECOMMENDED.

The LSP ingress is to send an MPLS echo request carrying the Extended BFD Discriminator TLV with the same BFD discriminator of the existing BFD session (one bootstrapped previously with the BFD Discriminator TLV), giving it an instance identifier. Once the transition of the existing BFD session is completed, then the LSP ingress can generate further MPLS echo request messages with the Extended BFD Discriminator TLV to bootstrap more BFD sessions.

7. Encapsulation

The encapsulation of BFD packets are the same as specified by [RFC5884]

8. Security Considerations

This document defines a mechanism to bootstrap multiple BFD sessions per FEC. BFD sessions, naturally, use system and network resources. More BFD sessions means more resources will be used. It is highly important to ensure only minimum number of BFD sessions are provisioned per FEC, and bootstrapped BFD sessions are properly deleted when no longer required. Additionally security measures described in [RFC4379] and [RFC5884] are to be followed.

9. IANA Considerations

9.1. Extended BFD Discriminator TLV

The IANA is requested to assign new value TBD1 for Extended BFD Discriminator TLV from the "Multiprotocol Label Switching Architecture (MPLS) Label Switched Paths (LSPs) Ping Parameters - TLVs" registry.

Value	Meaning	Reference
-----	-----	-----
TBD1	Extended BFD Discriminator TLV	this document

10. Acknowledgements

TBD

11. Contributing Authors

Girija Rao
Cisco Systems
Email: giraghav@cisco.com

Mallik Mudigonda
Cisco Systems
Email: mmudigon@cisco.com

12. 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.
- [RFC5880] Katz, D. and D. Ward, "Bidirectional Forwarding Detection (BFD)", RFC 5880, June 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.

Appendix A. Alternate format for the BFD Extended TLV

The BFD Extended TLV can be used to carry the Operation Type and the Operation Status (Op Status) bits that are defined below:

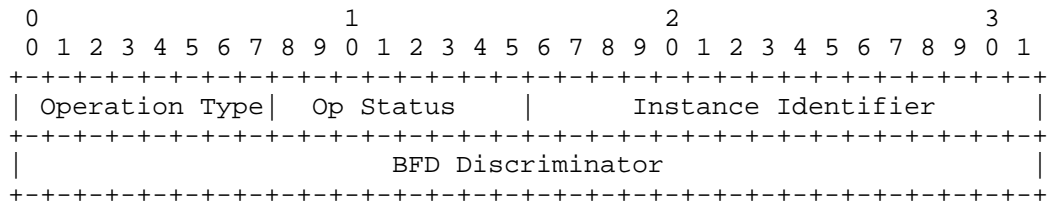


Figure 2: Alternate format of the Extended
BFD Discriminator TLV

Ed Note: The definitions of Operation Type and Operation Status fields are subject to discussion. Additional codes can be defined if this approach is pursued.

Operation Type - Operation to be performed on the corresponding BFD Discriminator. Valid values are:

- 1 - Create: This value MAY be used in the MPLS echo request, but MUST NOT be used in the MPLS echo reply. The operation type 1 indicates that receiver (i.e. LSP egress) is to ensure that BFD session for this FEC, with corresponding BFD Discriminator in "your discriminator" field, exists or is created.
- 2 - Delete: This value MAY be used in the MPLS echo request, but MUST NOT be used in the MPLS echo reply. The operation type 2 indicates that receiver (i.e. LSP egress) is to ensure that BFD session for this FEC, with corresponding BFD Discriminator in "your discriminator" field, does not exist or is deleted.
- 3 - CreateAck: This value MUST NOT be used in the MPLS echo request, but MAY be used in the MPLS echo reply. The operation type 3 indicates that receiver (i.e. LSP egress) is acknowledging received Create(1) request.
- 4 - DeleteAck: This value MUST NOT be used in the MPLS echo request, but MAY be used in the MPLS echo reply. The operation type 4 indicates that receiver (i.e. LSP egress) is acknowledging received Delete(2) request.

Op Status

- 0 - The operation succeeded.
- 1 - Not enough Resources.

BFD Discriminator - When the Extended BFD Discriminator TLV is carried in the MPLS echo request, this field describes the BFD discriminator allocated for this BFD session by the LSP ingress. When the Extended BFD Discriminator TLV is carried in the MPLS echo reply, this field describes the BFD discriminator allocated for this BFD session by the LSP egress.

The Extended BFD Discriminator TLV in an MPLS echo request MUST have either Create(1) or Delete(2) operation type. The Extended BFD Discriminator TLV in an MPLS echo reply MUST have either CreateAck(3) or DeleteACK(4) operation type.

Authors' Addresses

Vengada Prasad Govindan
Cisco Systems

Email: venggovi@cisco.com

Nobo Akiya
Cisco Systems

Email: nobo@cisco.com