

Internet Engineering Task Force
Internet-Draft
Intended status: Standards Track
Expires: January 17, 2013

H. Chen
Huawei Technologies
N. So
Tata Communications
A. Liu
Ericsson
L. Liu
KDDI R&D Lab Inc.
July 16, 2012

Extensions to RSVP-TE for P2MP LSP Egress Local Protection
draft-chen-mpls-p2mp-egress-protection-06.txt

Abstract

This document describes extensions to Resource Reservation Protocol - Traffic Engineering (RSVP-TE) for locally protecting egress nodes of a Traffic Engineered (TE) point-to-multipoint (P2MP) 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 17, 2013.

Copyright Notice

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

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

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

Table of Contents

1. Introduction	3
2. Terminology	4
3. Conventions Used in This Document	4
4. Mechanism	4
4.1. An Example of Egress Local Protection	4
4.2. Set up of Backup sub LSP	5
4.3. Forwarding State for Backup sub LSP(s)	6
4.4. Detection of Egress Node Failure	6
5. Egress Local Protection with FRR	7
6. Representation of a Backup Sub LSP	7
6.1. EGRESS_BACKUP_SUB_LSP Object	7
6.1.1. EGRESS_BACKUP_SUB_LSP IPv4 Object	8
6.1.2. EGRESS_BACKUP_SUB_LSP IPv6 Object	8
6.2. EGRESS_BACKUP_SECONDARY_EXPLICIT_ROUTE Object	9
7. Path Message	9
7.1. Format of Path Message	9
7.2. Processing of Path Message	10
8. Processing of Resv Message	11
9. IANA Considerations	11
10. Acknowledgement	11
11. References	12
11.1. Normative References	12
11.2. Informative References	12
Authors' Addresses	12

1. Introduction

RFC 4090 "Fast Reroute Extensions to RSVP-TE for LSP Tunnels" describes two methods for protecting P2P LSP tunnels or paths at local repair points. The first method is a one-to-one protection method, where a detour backup P2P LSP for each protected P2P LSP is created at each potential point of local repair, which is an intermediate node between the ingress node and the egress node of the protected LSP. The second method is a facility bypass backup protection method, where a bypass backup P2P LSP tunnel is created using MPLS label stacking to protect a potential failure point for a set of P2P LSP tunnels. The bypass backup tunnel can protect a set of P2P LSPs having similar backup constraints.

RFC 4875 "Extensions to RSVP-TE for P2MP TE LSPs" describes how to use the one-to-one protection method and facility bypass backup protection method to protect a link or intermediate node failure on the path of a P2MP LSP. However, there is no mention of locally protecting any egress node failure in a protected P2MP LSP.

An existing method for protecting the egress nodes of a P2MP LSP sets up a backup P2MP LSP from a backup ingress node to the backup egress nodes, where each egress node is paired with a backup egress node and protected by the backup egress node. The backup P2MP LSP carries the same traffic as the P2MP LSP at the same time. A traffic receiver from the P2MP LSP is normally connected to an egress node and its paired backup egress node. It receives the traffic from the egress node in normal situations.

The receiver selects the egress or backup egress node for receiving the traffic according to the route to the source through RPF. In a normal situation, it selects the egress node. When the egress node fails, it selects the backup egress for receiving the traffic since the route to the source through the egress node is gone and the route to the source through the backup egress node is active.

The main disadvantage of this method is that double network resources such as double bandwidths are used for protecting the egress nodes since the backup P2MP LSP consumes the same amount of network resource as the primary P2MP LSP. The impact on network efficiency can be significant in case of large P2MP deployments.

This document proposes a new method to locally protect the egress nodes of a P2MP LSP, which is called Egress Local Protection. It specifies the mechanism and extensions to RSVP-TE for locally protecting an egress node of a Traffic Engineered (TE) point-to-multipoint (P2MP) Label Switched Path through using a backup P2MP sub LSP. The new method overcomes the disadvantages described above.

The same extensions and mechanism can also be used to protect the egress node of a TE P2P LSP.

2. Terminology

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

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 RFC 2119.

4. Mechanism

This section briefly describes a solution that locally protects an egress node of a P2MP LSP through using a backup P2MP sub LSP. We first show an example, and then present different parts of the solution, which includes the creation of the backup sub LSP, the forwarding state for the backup sub LSP, and the detection of a failure in the egress node.

4.1. An Example of Egress Local Protection

Figure 1 below illustrates an example of using backup sub LSPs to locally protect egress nodes of a P2MP LSP. The P2MP LSP is from ingress node R1 to three egress nodes: L1, L2 and L3. It is represented by double lines in the figure.

La, Lb and Lc are the designated backup egress nodes for the egress nodes L1, L2 and L3 of the P2MP LSP respectively. In order to distinguish an egress node (e.g., L1 in the figure) and a backup egress node (e.g., La in the figure), an egress node is called a primary egress node in the following description.

The backup sub LSP used to protect the primary egress node L1 is from its previous hop node R3 to the backup egress node La. The backup sub LSP used to protect the primary egress node L2 is from its previous hop node R5 to the backup egress node Lb. The backup sub LSP used to protect the primary egress node L3 is from its previous hop node R5 to the backup egress node Lc via the intermediate node Rc.

During normal operation, the traffic transported by the P2MP LSP is

forwarded through R3 to L1, then delivered to its destination CE1. When the failure of L1 is detected, R3 forwards the traffic to the backup egress node La, which then delivers the traffic to its destination CE1. The time for switching the traffic after L1 fails is within tens of milliseconds.

L1's failure CAN be detected by a BFD session between L1 and R3.

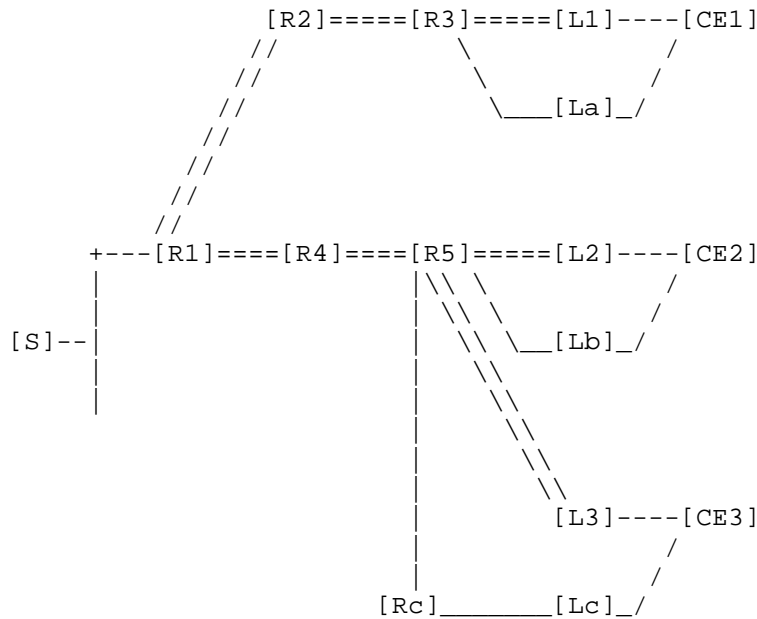


Figure 1: P2MP sub LSP for Locally Protecting Egress

4.2. Set up of Backup sub LSP

A backup egress node is designated for a primary egress node of a LSP. The previous hop node of the primary egress node sets up a backup sub LSP from itself to the backup egress node after receiving the information about the backup egress node.

The previous hop node sets up the backup sub LSP, creates and maintains its state in the same way as of setting up a source to leaf (S2L) sub LSP from the signalling's point of view. It constructs and sends a RSVP-TE PATH message along the path for the backup sub LSP, receives and processes a RSVP-TE RESV message that responds to the PATH message.

4.3. Forwarding State for Backup sub LSP(s)

The forwarding state for the backup sub LSP is different from that for a P2MP S2L sub LSP. After receiving the RSVP-TE RESV message for the backup sub LSP, the previous hop node 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. It SHALL only be used after the failure of the primary egress node.

Upon detection of the primary egress node failure, the state or flag of the forwarding entry for the backup sub LSP is set to be active. Thus, the previous hop node of the primary egress node will forward the traffic to the backup egress node through the backup sub LSP, which then send the traffic to its destination.

4.4. Detection of Egress Node Failure

The previous hop node of the primary egress node SHALL detect four types of failures described below:

- o The failure of the primary egress node (e.g. L1 in Figure 1)
- o The failure of the link between the primary egress node and its previous hop node (e.g. the link between R3 and L1 in Figure 1)
- o The failure of the destination node for the primary egress node (e.g. CE1 in Figure 1)
- o The failure of the link between the primary egress node and its destination node (e.g. the failure of the link between L1 and CE1 in Figure 1).

Failure of the primary egress node and the link between itself and its previous hop node CAN be detected through a BFD session between itself and its previous hop node in MPLS networks.

In the GMPLS networks where the control plane and data plane are physically separated, the detection and localization of failures in the physical layer can be achieved by introducing the link management protocol (LMP) or assisting by performance monitoring devices.

Failure of the destination node and the link between the primary egress node and the destination node CAN be detected by a BFD session between the previous hop node and the destination node.

Upon detecting any above mentioned failures, the previous hop node imports the traffic from the LSP into the backup sub LSP. The

traffic is then delivered to its destination through the backup egress node.

5. Egress Local Protection with FRR

RFC4875 "Extensions to RSVP-TE for P2MP TE LSPs" describes how to use RFC 4090 "Fast Reroute Extensions to RSVP-TE for LSP Tunnels" (FRR for short) to locally protect failures in a link or intermediate node of a P2MP LSP. However, there is not any standard that locally protects the egresses of the P2MP LSP. The egress local protection mechanism proposed in this document fills this gap. Thus, through using the egress local protection and the FRR, we can locally protect the egress nodes, all the links and the intermediate nodes of a P2MP LSP. The traffic switchover time is within tens of milliseconds whenever any of the egresses, the links and the intermediate nodes of the P2MP LSP fails.

All the egress nodes of the P2MP LSP can be locally protected through using the egress local protection. All the links and the intermediate nodes of the LSP can be locally protected by using the FRR. Note that the methods for locally protecting all the links and the intermediate nodes of a P2MP LSP are out of scope of this document.

6. Representation of a Backup Sub LSP

A backup sub LSP exists within the context of a P2MP LSP in a way similar to a S2L sub LSP. It is identified by the P2MP LSP ID, Tunnel ID, and Extended Tunnel ID in the SESSION object, the tunnel sender address and LSP ID in the SENDER_TEMPLATE object, and the backup sub LSP destination address in the EGRESS_BACKUP_SUB_LSP object (to be defined in the section below).

An EGRESS_BACKUP_SECONDARY_EXPLICIT_ROUTE Object (EB-SERO) is used to optionally specify the explicit route of a backup sub LSP that is from a previous hop node to a backup egress node. The EB-SERO is defined in the following section.

6.1. EGRESS_BACKUP_SUB_LSP Object

An EGRESS_BACKUP_SUB_LSP object identifies a particular backup sub LSP belonging to the LSP.

6.1.1.1. EGRESS_BACKUP_SUB_LSP IPv4 Object

The class of the EGRESS_BACKUP_SUB_LSP IPv4 object is the same as that of the S2L_SUB_LSP IPv4 object defined in RFC 4875. The C-Type of the object is a new number 3, or may be another number assigned by Internet Assigned Numbers Authority (IANA).

EGRESS_BACKUP_SUB_LSP Class = 50,
EGRESS_BACKUP_SUB_LSP_IPv4 C-Type = 3

```

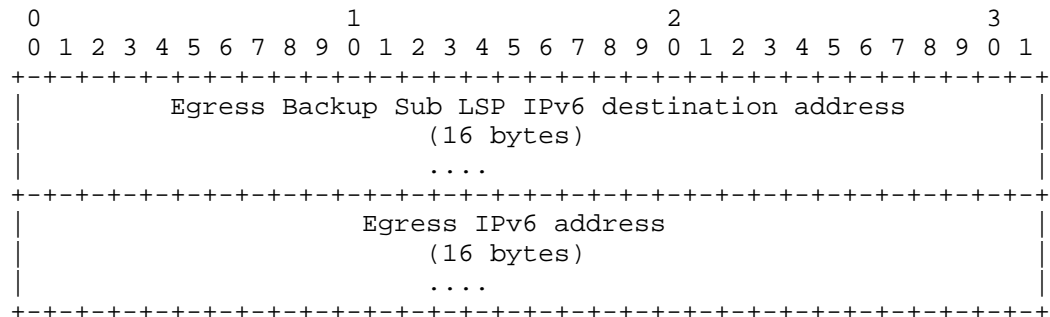
      0               1               2               3
    0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-----+-----+-----+-----+-----+-----+-----+-----+
|               Egress Backup Sub LSP IPv4 destination address               |
+-----+-----+-----+-----+-----+-----+-----+-----+
|               Egress IPv4 address               |
+-----+-----+-----+-----+-----+-----+-----+-----+

```

Egress Backup Sub LSP IPv4 destination address
IPv4 address of the backup sub LSP destination is the backup egress node.
Egress IPv4 address
IPv4 address of the egress node

6.1.1.2. EGRESS_BACKUP_SUB_LSP IPv6 Object

The class of the EGRESS_BACKUP_SUB_LSP IPv6 object is the same as that of the S2L_SUB_LSP IPv6 object defined in RFC 4875. The C-Type of the object is a new number 4, or may be another number assigned by Internet Assigned Numbers Authority (IANA).



```

<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 sub LSP descriptor list>]

```

The format of the egress backup sub LSP descriptor list in the enhanced Path message is defined as follows.

```

<egress backup sub LSP descriptor list> ::=
    <egress backup sub LSP descriptor>
    [ <egress backup sub LSP descriptor list> ]

<egress backup sub LSP descriptor> ::=
    <EGRESS_BACKUP_SUB_LSP>
    [ <EGRESS_BACKUP_SECONDARY_EXPLICIT_ROUTE> ]

```

7.2. Processing of Path Message

The ingress node of a LSP initiates a Path message with an egress backup sub LSP descriptor list for protecting primary egress nodes of the LSP. In order to protect a primary egress node of the LSP, the ingress node MUST add an EGRESS_BACKUP_SUB_LSP object into the list. The object contains the information about the backup egress node to be used to protect the failure of the primary egress node. An EGRESS_BACKUP_SECONDARY_EXPLICIT_ROUTE object (EB-SERO), which describes an explicit path to the backup egress node, SHALL follow the EGRESS_BACKUP_SUB_LSP.

If the previous hop node of the primary egress node receives the Path message with an egress backup sub LSP descriptor list, it generates a new Path message based on the information in the EGRESS_BACKUP_SUB_LSP (and according to EB-SERO if it exists) containing the backup egress node.

The format of this new Path message is the same as that of the Path message defined in RFC 4875. This new Path message is used to signal the segment of a special S2L sub-LSP from the previous hop node to the backup egress node. The new Path message is sent to the next-hop node along the path for the backup sub LSP.

If an intermediate node receives the Path message with an egress backup sub LSP descriptor list. Then it MUST put the EGRESS_BACKUP_SUB_LSP (according to EB-SERO if exists) containing a backup egress into a Path message to be sent towards the backup egress. This SHALL be done for each EGRESS_BACKUP_SUB_LSP containing a backup egress node in the list.

When a primary egress node of the LSP receives the Path message with an egress backup sub LSP descriptor list, it SHOULD ignore the egress backup sub LSP descriptor list and generate a PathErr message.

8. Processing of Resv Message

The format of the Resv Message is not changed. The processing of the Resv Message at the previous hop of a primary egress node is enhanced for reporting the status of the primary egress protection.

The previous hop node of the primary egress node sets the protection flags in the RRO IPv4/IPv6 Sub-object for the primary egress node according to the status of the primary egress node and the backup sub LSP protecting the primary egress node. For example, it will set the node protection bit to one indicating that the primary egress node is protected when the backup sub LSP to the backup egress node is set up for protecting the primary egress node. It will set the bandwidth protection bit to one when the backup sub LSP guarantees to provide the desired bandwidth that is specified in the FAST_REROUTE object or the bandwidth of the protected LSP.

9. IANA Considerations

TBD

10. Acknowledgement

The authors would like to thank Richard Li, Olufemi Komolafe, 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.
- [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: Huaimochen@huawei.com

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

Email: ning.so@tatacommunications.com

Autumn Liu
Ericsson
CA
USA

Email: autumn.liu@ericsson.com

Lei Liu
KDDI R&D Lab Inc.
2-1-15
Ohara Fujimino-shi, Saitama
Japan

Email: le-liu@kddilabs.jp

Internet Engineering Task Force
Internet-Draft
Intended status: Standards Track
Expires: January 17, 2013

H. Chen
Huawei Technologies
N. So
Tata Communications
A. Liu
Ericsson
L. Liu
KDDI R&D Lab Inc.
July 16, 2012

Extensions to RSVP-TE for P2MP LSP Ingress Local Protection
draft-chen-mpls-p2mp-ingress-protection-06.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) Point-to-MultiPoint (P2MP) 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 17, 2013.

Copyright Notice

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

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

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

Table of Contents

1. Introduction	3
2. Terminology	4
3. Conventions Used in This Document	4
4. Mechanism	4
4.1. An Example of Ingress Local Protection	4
4.2. Set up of Backup P2MP sub Tree	5
4.3. Forwarding State for Backup P2MP sub Tree	5
4.4. Detection of Failure around Ingress	6
5. Ingress Local Protection with FRR	7
6. LSP Information Message	7
6.1. Format of LSP Information Message	8
6.2. Processing of LSP Information Message	8
6.3. Discussions on Other Approaches	9
7. LSP Information Confirmation Message	9
7.1. Format of LSP Information Confirmation Message	9
7.2. Processing of LSP Information Confirmation Message	10
8. PATH Messages for Backup P2MP sub Tree	10
8.1. Construction of PATH Messages	10
8.2. Processing of PATH Messages	11
9. IANA Considerations	11
10. Acknowledgement	11
11. References	12
11.1. Normative References	12
11.2. Informative References	12
Authors' Addresses	13

1. Introduction

RFC4090 "Fast Reroute Extensions to RSVP-TE for LSP Tunnels" describes two methods to protect P2P LSP tunnels or paths at local repair points. The first method is a one-to-one backup method, where a detour backup P2P LSP for each protected P2P LSP is created at each potential point of local repair, which is an intermediate node between the ingress node and the egress node of the protected LSP. The second method is a facility bypass backup protection method, where a bypass backup P2P LSP tunnel is created using MPLS label stacking to protect a potential failure point for a set of P2P LSP tunnels. The bypass backup tunnel can protect a set of P2P LSPs that have similar backup constraints.

RFC4875 "Extensions to RSVP-TE for P2MP TE LSPs" describes how to use the one-to-one backup method and facility bypass backup method to protect a link or intermediate node failure on the path of a P2MP LSP. However, there is no mention of locally protecting an ingress node failure in a protected P2MP LSP.

There exist two methods for protecting an ingress node of a P2MP LSP. The first method deploys a backup P2MP LSP from a backup ingress node to the destination nodes to protect the ingress node. The main disadvantage of this method is that the backup P2MP LSP consumes additional network bandwidth along the entire LSP paths. The impact on network efficiency can be significant in case of large P2MP deployments. In addition, the backup LSP often has to be manually constructed so that the backup P2MP LSP does not route through the unprotected ingress node, and it has to be linked to the primary LSP logically at the head-end to allow the fast switching in case of ingress failure.

The second method extends the existing ways of protecting an intermediate node of a P2P LSP to protect an ingress node of a P2MP LSP. The disadvantages of this method include extra work for refreshing PATH messages and processing RESV messages for the P2MP LSP in the backup ingress node.

This document defines extensions to RSVP-TE for locally protecting an ingress node of a Traffic Engineered (TE) point-to-multipoint (P2MP) Label Switched Path (LSP) through using a backup P2MP sub tree. The new method overcomes the disadvantages described above. It can also be applied for protecting an ingress node of a TE point-to-point (P2P) LSP since a TE P2P LSP can be considered as a special case of a TE P2MP LSP.

2. Terminology

This document uses terminologies defined in RFC2205, RFC3031, RFC3209, RFC3473, RFC4090, RFC4461, and RFC4875.

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 RFC 2119.

4. Mechanism

This section briefly describes a solution that locally protects an ingress node of a P2MP LSP through using a backup P2MP sub tree. We start with a simple example, and then present different parts of the solution, which includes the creation of the backup P2MP sub tree, the forwarding state for the backup P2MP subtree, and the detection of a failure in the ingress node.

4.1. An Example of Ingress Local Protection

Figure 1 below illustrates an example of using a backup P2MP sub tree to locally protect the ingress of a P2MP LSP. The P2MP LSP to be protected is from ingress node R1 to three egress/leaf nodes: L1, L2 and L3. The backup P2MP sub tree used to protect the ingress node R1 is from backup ingress node Ra to the next hop nodes R2 and R4 of the ingress node R1 along the P2MP LSP.

The traffic from source S may be delivered to both R1 and Ra. R1 introduces the traffic into the P2MP LSP, which is sent to the egress/leaf nodes L1, L2 and L3 along the P2MP LSP. Ra normally does not put the traffic into the backup P2MP sub tree, which is from Ra to R2 and R4.

There may be a BFD session between ingress node R1 and backup ingress node Ra. Ra uses this BFD session to detect the failure of ingress R1. When Ra detects the failure of R1, it imports the traffic from the source S into the backup P2MP sub tree. The traffic from the sub tree is merged into the P2MP LSP at R2 and R4, and then sent to the egress/leaf nodes L1, L2 and L3 along the P2MP LSP. The time for switching the traffic after R1 fails is within tens of milliseconds.

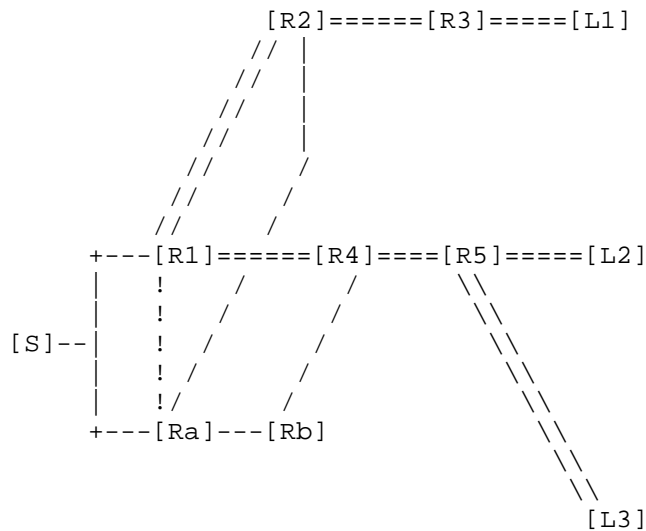


Figure 1: P2MP sub Tree for Locally Protecting Ingress

After the failure of the ingress node R1, the refresh of the PATH messages for the ingress node is not needed. Each of the next-hop nodes of the ingress node will receive the PATH messages and the refresh of the PATH messages for the backup P2MP sub tree from the backup ingress node Ra, which make the P2MP LSP alive.

4.2. Set up of Backup P2MP sub Tree

For the ingress node of the P2MP LSP, a backup ingress node is designated to protect it. The ingress node sends the P2MP LSP information to the backup ingress node. The backup ingress node initiates the creation of the backup P2MP sub tree from itself to the next-hop nodes of the ingress node.

The backup ingress node sets up the backup P2MP sub tree in a way similar to setting up a P2MP tree or LSP from the signaling's point of view. It constructs and sends RSVP-TE PATH messages along the path for the backup P2MP sub tree with the final destinations (i.e, egress/leaf nodes) matching the P2MP LSP. It receives and processes RSVP-TE RESV messages that response to the PATH messages.

4.3. Forwarding State for Backup P2MP sub Tree

The forwarding state for the backup P2MP sub tree is different from that for a P2MP LSP. After receiving the RSVP-TE RESV messages for the backup P2MP sub tree, the backup ingress node creates a

forwarding entry with an inactive state or flag. This forwarding entry with an inactive state or flag is called an inactive forwarding entry. In a normal operation, this inactive forwarding entry is not used to forward any data traffic to be transported by the P2MP LSP, even though the data traffic may be delivered to the backup ingress node from an external node such as source node S in the above example or network. The forwarding entry for the P2MP LSP is with an active state or flag. Thus when the data traffic from the external node or network reaches the ingress node of the P2MP LSP, it is imported into the P2MP LSP tunnel through the active forwarding entry on the ingress node.

When the ingress node fails, the inactive forwarding entry on the backup ingress node is changed to active. Thus when the data traffic from the external node reaches the backup ingress node, it is imported into the backup P2MP sub tree. When the traffic arrives at the next-hop nodes through the backup P2MP sub tree, it is merged into the P2MP LSP to be transported to the destinations.

4.4. Detection of Failure around Ingress

There can be two different failure scenarios involving the ingress node of a P2MP LSP that need to be detected.

- o The failure of the ingress node (e.g. R1 of figure 1).
- o The failure of the link between the source node and the ingress node (e.g. the link between node S and node R1 in figure 1).

A failure of the ingress node can be detected through a BFD session between the ingress node and the backup ingress node in MPLS networks. A failure of the link between the source node and the ingress node can be detected by a BFD session running over the link and to the backup ingress via the ingress.

In the GMPLS networks where the control plane and data plane are physically separated, the detection and localization of failures in the physical layer can be achieved by introducing the link management protocol (LMP) or assisting by performance monitoring devices.

After the backup ingress node detects any failure involving the ingress node, it imports the traffic from the source node into the backup P2MP sub tree. The traffic from the backup ingress node via the sub tree is merged into the P2MP LSP on the next-hop nodes of the ingress of the P2MP LSP, and then transported to the egress/leaf nodes of the P2MP LSP.

5. Ingress Local Protection with FRR

RFC4875 "Extensions to RSVP-TE for P2MP TE LSPs" describes how to use RFC 4090 "Fast Reroute Extensions to RSVP-TE for LSP Tunnels" (FRR for short) to locally protect failures in a link or intermediate node of a P2MP LSP. However, there is not any standard that locally protects the ingress of the P2MP LSP. The ingress local protection mechanism described above fills this gap. Thus, through using the ingress local protection and the FRR, we can locally protect the ingress node, all the links and the intermediate nodes of a P2MP LSP. The traffic switchover time is within tens of milliseconds whenever the ingress, any of the links and the intermediate nodes of the P2MP LSP fails.

The ingress node of the P2MP LSP can be locally protected through using the ingress local protection. All the links and all the intermediate nodes of the P2MP LSP can be locally protected through using the FRR.

RFC 4090 defines fast reroute extensions to RSVP-TE for local protection of P2P TE LSP in MPLS networks. RFC 4090, which is for local protection of P2P TE LSP, has a few of limitations or issues when it is used for local protection of P2MP TE LSP.

For example, locally protecting an intermediate node of a P2MP TE LSP requires, when the protected node is a branch LSR, a set of P2P Next-Next-Hop (NNHOP) Bypass tunnels toward all LSRs downstream to the protected node. When the protected node fails, the PLR has to replicate traffic on each of the P2P bypass tunnels. If there are K next-next-hops, this may lead to K times of the traffic on some links, which is not acceptable.

To overcome these limitations, draft "P2MP MPLS-TE Fast Reroute with P2MP Bypass Tunnels" proposes extensions to FRR procedures defined in RFC4090 to locally protect links and intermediate nodes of a P2MP TE LSP with P2MP bypass tunnels.

Note that the methods for locally protecting all the links and the intermediate nodes of a P2MP LSP are out of scope of this document.

6. LSP Information Message

LSP information messages are used to transfer the information about a P2MP LSP to a backup ingress node from an ingress node. The destination address of the LSP information message is that of the backup ingress node. This section describes the format of an LSP information message and processing of the message. It also discusses

other approaches for transferring the information about a P2MP LSP to a backup ingress from an ingress.

6.1. Format of LSP Information Message

The format of a P2MP LSP information message is illustrated below.

```
<LSP Information 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>]
    <RECORD_ROUTE>
    <S2L sub LSP flow descriptor list>
```

The formats and values of the objects in a P2MP LSP information message are similar to or the same as those of the corresponding objects defined in RFC4875.

The value of the Msg Type field in the common header in the P2MP LSP information message will be a new number to be assigned by Internet Assigned Numbers Authority (IANA).

6.2. Processing of LSP Information Message

Similar to sending an existing RSVP-TE message such as a PATH message, the primary ingress MUST send a updated RSVP-TE LSP information message to the backup ingress whenever there is a change in the RSVP-TE LSP information message. It MAY send the same RSVP-TE LSP information message to the backup ingress every refresh interval if there is no change.

When the backup ingress receives the RSVP-TE LSP information message from the primary ingress, it stores the LSP information, constructs PATH messages, and sends the PATH messages downstream accordingly.

If it has not received any RSVP-TE LSP information message for an extended period of time (e.g. a cleanup timeout interval) and the BFD session between the primary ingress and backup ingress is up, it SHALL remove the information about the P2MP LSP, constructs PathTear messages, and send the PathTear messages downstream accordingly.

When the BFD session between the primary ingress and backup ingress is down, the backup ingress MUST keep the information about the P2MP LSP and the state of the backup P2MP sub tree even though it has not received any RSVP-TE LSP information message for an extended period of time. It refreshes the PATH messages downstream for the backup P2MP sub tree.

6.3. Discussions on Other Approaches

The information about a P2MP LSP may be transferred through other approaches from the ingress node of the LSP to the backup ingress node. One approach is to use OSPF Opaque LSA. The main reason for giving up this option is that more parts need to be changed. Both OSPF and RSVP-TE need to be modified.

On the ingress node, RSVP-TE needs to be changed to send the information to OSPF when there is a change on the information about the P2MP LSP. OSPF needs to be changed to receive the information about the P2MP LSP from RSVP-TE and distribute the information in Opaque LSA to the OSPF on the backup ingress node.

On the backup ingress node, OSPF needs to be changed to receive the information in Opaque LSA from the ingress node and send the information to RSVP-TE. RSVP-TE needs to be changed to receive the information about the P2MP LSP from OSPF.

7. LSP Information Confirmation Message

LSP information confirmation messages are used to confirm that the corresponding LSP information messages are received. With the confirmation messages, the refresh of the LSP information messages is not needed. In addition, the state of the backup P2MP sub tree and the action of switching over of traffic are communicated with the primary ingress through the messages. This section describes the format of an LSP information confirmation message and processing of the message.

7.1. Format of LSP Information Confirmation Message

The format of a P2MP LSP information confirmation message is illustrated below.

```
<LSP Information Confirmation Message> ::=
    <Common Header> [ <INTEGRITY> ]
    [ [ <MESSAGE_ID_ACK> | <MESSAGE_ID_NACK> ] ... ]
    [ <MESSAGE_ID> ]
    <SESSION> <RSVP_HOP>
    <sender descriptor>
```

The formats and values of the objects in a P2MP LSP information confirmation message are similar to or the same as those of the corresponding objects defined in RFC4875.

The value of the Msg Type field in the common header in the P2MP LSP information confirmation message will be a new number such as 69 for the LSP information confirmation message, or may be another number assigned by Internet Assigned Numbers Authority (IANA).

7.2. Processing of LSP Information Confirmation Message

When the backup ingress node receives a RSVP-TE LSP information message from the ingress node, it SHALL construct and send an LSP confirmation message to the ingress node to acknowledge the message received.

After the ingress node receives the LSP confirmation message, it SHOULD stop refreshing the LSP information message.

8. PATH Messages for Backup P2MP sub Tree

PATH messages for a backup P2MP sub tree has the same format as PATH messages for a P2MP LSP defined in RFC 4875. This section describes the construction of the PATH messages for the backup P2MP sub tree, which is followed by processing of the PATH messages.

8.1. Construction of PATH Messages

When the backup ingress node receives a P2MP LSP information message, it checks to see if anything has been changed. If the message is a new message or the information in the message has been changed, then the PATH messages for the backup P2MP sub tree are to be constructed as follows.

First, a path to the next-hop nodes of the ingress node HAS to be computed. The path MUST satisfy the constraints for the P2MP LSP and not go through the ingress node.

If a path is computed successfully, then the PATH messages for the

backup P2MP sub tree are constructed based on the computed path and the information message received, and sent downstream accordingly. After sending the PATH messages, the backup ingress node receives RESV messages from downstream nodes responding to the PATH messages. It then processes the RESV messages and creates forwarding state based on the information in the RESV messages.

If a path can not be found, the backup ingress node SHALL tear down the backup P2MP sub tree created based the previous information message.

The construction of a PATH message on a backup ingress node for a backup P2MP sub tree is similar to the construction of a normal PATH message on an ingress node for a P2MP LSP. It is based on LSP information messages and a computed path for the backup P2MP sub tree. The backup ingress node refreshes the PATH message to its downstream nodes when the refresh reduction is not enabled.

The EXPLICIT_ROUTE object and the objects in the S2L sub-LSP descriptor list for the PATH message may be constructed through combining the path computed to the next-hop nodes of the ingress node and the path from the next-hop nodes to the destination nodes of the P2MP LSP obtained from the RECORD_ROUTE object and the objects for the S2L sub-LSP flow descriptor list in the LSP information messages.

8.2. Processing of PATH Messages

The processing of PATH messages on the intermediate nodes and the destination nodes along the backup P2MP sub tree is the same as the processing of PATH messages for a P2MP LSP.

9. IANA Considerations

TBD

10. Acknowledgement

The authors would like to thank Richard Li, Rahul Aggarwal, 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.
- [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.

Authors' Addresses

Huaimo Chen
Huawei Technologies
Boston, MA
USA

Email: Huaimochen@huawei.com

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

Email: ning.so@tatacommunications.com

Autumn Liu
Ericsson
CA
USA

Email: autumn.liu@ericsson.com

Lei Liu
KDDI R&D Lab Inc.
2-1-15
Ohara Fujimino-shi, Saitama
Japan

Email: le-liu@kddilabs.jp

MPLS Working Group
Internet-Draft
Intended status: Informational
Expires: February 3, 2013

D. Frost, Ed.
S. Bryant, Ed.
Cisco Systems
M. Bocci, Ed.
Alcatel-Lucent
L. Berger, Ed.
LabN Consulting
August 2, 2012

A Framework for Point-to-Multipoint MPLS in Transport Networks
draft-fbb-mpls-tp-p2mp-framework-05

Abstract

The Multiprotocol Label Switching (MPLS) Transport Profile (MPLS-TP) is the common set of MPLS protocol functions defined to enable the construction and operation of packet transport networks. The MPLS-TP supports both point-to-point and point-to-multipoint transport paths. This document defines the elements and functions of the MPLS-TP architecture applicable specifically to supporting point-to-multipoint transport paths.

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

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 February 3, 2013.

Copyright Notice

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

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

Table of Contents

1. Introduction	3
1.1. Scope	3
1.2. Terminology	4
1.2.1. Additional Definitions and Terminology	4
1.3. Applicability	4
2. MPLS Transport Profile Point-to-Multipoint Requirements . . .	4
3. Architecture	5
3.1. MPLS-TP Encapsulation and Forwarding	6
4. Operations, Administration and Maintenance (OAM)	6
5. Control Plane	6
5.1. Point-to-Multipoint LSP Control Plane	7
5.2. Point-to-Multipoint PW Control Plane	7
6. Survivability	8
7. Network Management	8
8. Security Considerations	8
9. IANA Considerations	9
10. References	9
10.1. Normative References	9
10.2. Informative References	10

1. Introduction

The Multiprotocol Label Switching (MPLS) Transport Profile (MPLS-TP) is the common set of MPLS protocol functions defined to meet the requirements specified in [RFC5654]. The MPLS-TP Framework [RFC5921] provides an overall introduction to the MPLS-TP and defines the general architecture of the Transport Profile, as well as those aspects specific to point-to-point transport paths. The purpose of this document is to define the elements and functions of the MPLS-TP architecture applicable specifically to supporting point-to-multipoint transport paths.

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

1.1. Scope

This document defines the elements and functions of the MPLS-TP architecture related to supporting point-to-multipoint transport paths. The reader is referred to [RFC5921] for those aspects of the MPLS-TP architecture that are generic, or concerned specifically with point-to-point transport paths.

1.2. Terminology

Term	Definition
-----	-----
LSP	Label Switched Path
MPLS-TP	MPLS Transport Profile
SDH	Synchronous Digital Hierarchy
ATM	Asynchronous Transfer Mode
OTN	Optical Transport Network
OAM	Operations, Administration and Maintenance
G-ACh	Generic Associated Channel
GAL	G-ACh Label
MEP	Maintenance End Point
MIP	Maintenance Intermediate Point
APS	Automatic Protection Switching
SCC	Signaling Communication Channel
MCC	Management Communication Channel
EMF	Equipment Management Function
FM	Fault Management
CM	Configuration Management
PM	Performance Management
LSR	Label Switching Router
MPLS-TE	MPLS Traffic Engineering
P2MP	Point-to-multipoint
PW	Pseudowire

1.2.1. Additional Definitions and Terminology

Detailed definitions and additional terminology may be found in [RFC5921] and [RFC5654].

1.3. Applicability

The point-to-multipoint connectivity provided by an MPLS-TP network is based on the point-to-multipoint connectivity provided by MPLS networks. MPLS TE-LSP support is discussed in [RFC4875] and [RFC5332], and PW support is being developed based on [I-D.ietf-pwe3-p2mp-pw-requirements] and [I-D.ietf-l2vpn-vpms-frmwk-requirements]. MPLS-TP point-to-multipoint connectivity is analogous to that provided by traditional transport technologies such as Optical Transport Network (OTN) point-to-multipoint [ref?] and optical drop-and-continue [ref?], and thus supports the same class of traditional applications.

2. MPLS Transport Profile Point-to-Multipoint Requirements

The requirements for MPLS-TP are specified in [RFC5654], [RFC5860], and [RFC5951]. This section provides a brief summary of point-to-

multipoint transport requirements as set out in those documents; the reader is referred to the documents themselves for the definitive and complete list of requirements.

- o MPLS-TP must support unidirectional point-to-multipoint (P2MP) transport paths.
- o MPLS-TP must support traffic-engineered point-to-multipoint transport paths.
- o MPLS-TP must be capable of using P2MP server (sub)layer capabilities as well as P2P server (sub)layer capabilities when supporting P2MP MPLS-TP transport paths.
- o The MPLS-TP control plane must support establishing all the connectivity patterns defined for the MPLS-TP data plane (i.e., unidirectional P2P, associated bidirectional P2P, co-routed bidirectional P2P, unidirectional P2MP) including configuration of protection functions and any associated maintenance functions.
- o Recovery techniques used for P2P and P2MP should be identical to simplify implementation and operation.
- o Unidirectional 1+1 and 1:n protection for P2MP connectivity must be supported.
- o MPLS-TP recovery in a ring must protect unidirectional P2MP transport paths.

3. Architecture

The overall architecture of the MPLS Transport Profile is defined in [RFC5921]. The architecture for point-to-multipoint MPLS-TP comprises the following additional elements and functions:

- o Unidirectional point-to-multipoint Label Switched Paths (LSPs)
- o Unidirectional point-to-multipoint pseudowires (PWs)
- o Optional point-to-multipoint LSP and PW control planes
- o Survivability, network management, and Operations, Administration and Maintenance (OAM) functions for point-to-multipoint PWs and LSPs

The following subsections summarise the encapsulation and forwarding of point-to-multipoint traffic within an MPLS-TP network, and the encapsulation options for delivery of traffic to and from MPLS-TP

Customer Edge devices when the network is providing a packet transport service.

3.1. MPLS-TP Encapsulation and Forwarding

Packet encapsulation and forwarding for MPLS-TP point-to-multipoint LSPs is identical to that for MPLS-TE point-to-multipoint LSPs. MPLS-TE point-to-multipoint LSPs were introduced in [RFC4875] and the related data-plane behaviour was further clarified in [RFC5332]. MPLS-TP allows for both upstream-assigned and downstream-assigned labels for use with point-to-multipoint LSPs.

Packet encapsulation and forwarding for point-to-multipoint PWs is currently being defined by the PWE3 Working Group [I-D.raggarwa-pwe3-p2mp-pw-encaps].

4. Operations, Administration and Maintenance (OAM)

The overall OAM architecture for MPLS-TP is defined in [RFC6371], and P2MP OAM design considerations are described in Section 3.7 of that RFC.

All the traffic sent over a P2MP transport path, including OAM packets generated by a MEP, is sent (multicast) from the root to all the leaves, thus every OAM packet is sent to all leaves, and thus can simultaneously instrument all the MEs in a P2MP MEG. If an OAM packet is to be processed by only one leaf, it requires information to indicate to all other leaves that the packet must be discarded. To address a packet to an intermediate node in the tree, TTL based addressing is used to set the radius and addressing information in the OAM payload is used to identify the specific destination node.

P2MP paths are unidirectional; therefore, any return path to an originating MEP for on-demand transactions will be out-of-band. Out of band return paths are discussed in Section 3.8 of [RFC5921]

[Editor's note: Additional information / text has been published in [I-D.hmk-mpls-tp-p2mp-oam-framework]. The Editors will coordinate with the draft authors to identify which text should be folded into this document and which should remain in a standalone document.]

5. Control Plane

The framework for the MPLS-TP control plane is provided in [RFC6373]. This document reviews MPLS-TP control plane requirements as well as provides details on how the MPLS-TP control plane satisfies these requirements. Most of the requirements identified in [RFC6373] apply equally to P2P and P2MP transport paths. The key P2MP specific

control plane requirements are identified in requirement 6 (P2MP transport paths), 34 (use P2P sub-layers), 49 (common recovery solutions for P2P and P2MP), 59 (1+1 protection), 62 (1:n protection), and 65 (1:n shared mesh recovery).

[RFC6373] defines the control plane approach used to support MPLS-TP transport paths. It identifies Generalized MPLS (GMPLS) as the control plane for MPLS-TP Label Switched Paths (LSPs) and Targeted LDP (T-LDP) as the control plane for pseudowires (PWs). MPLS-TP allows that either, or both, LSPs and PWs to be provisioned statically or via a control plane. As noted in [RFC6373]:

The PW and LSP control planes, collectively, must satisfy the MPLS-TP control-plane requirements. As with P2P services, when P2MP client services are provided directly via LSPs, all requirements must be satisfied by the LSP control plane. When client services are provided via PWs, the PW and LSP control planes can operate in combination, and some functions may be satisfied via the PW control plane while others are provided to PWs by the LSP control plane. This is particularly noteworthy for P2MP recovery.

5.1. Point-to-Multipoint LSP Control Plane

The MPLS-TP control plane for point-to-multipoint LSPs uses GMPLS and is based on Resource Reservation Protocol - Traffic Engineering (RSVP-TE) for point-to-multipoint LSPs as defined in [RFC4875]. A detailed listing of how GMPLS satisfies MPLS-TP control plane requirements is provided in [RFC6373].

Per [RFC6373], the definitions of P2MP, [RFC4875], and GMPLS recovery, [RFC4872] and [RFC4873], do not explicitly cover their interactions. MPLS-TP requires a formal definition of recovery techniques for P2MP LSPs. Such a formal definition will be based on existing RFCs and may not require any new protocol mechanisms but, nonetheless, should be documented. Protection of P2MP LSPs is also discussed in [RFC6372] Section 4.7.3.

5.2. Point-to-Multipoint PW Control Plane

[I-D.ietf-pwe3-p2mp-pw] The MPLS-TP control plane for point-to-multipoint PWs uses the LDP P2MP signaling extensions for PWs defined in [I-D.ietf-pwe3-p2mp-pw]. This definition is limited to single segment PWs and is based on LDP [RFC5036] with upstream-assigned labels [RFC5331]. The document does not address recovery of P2MP PWs. Such recovery can be provided via P2MP LSP recovery as generally discussed in [RFC6372]. Alternatively, PW recovery [I-D.ietf-pwe3-redundancy] can be extended to explicitly support recovery of P2MP PWs.

6. Survivability

The overall survivability architecture for MPLS-TP is defined in [RFC6372], and section 4.7.3 in particular describes the application of linear protection to unidirectional P2MP entities using 1+1 protection architecture. The approach is for the root of the P2MP tree to bridge the user traffic to both the working and protection entities. Each sink/leaf MPLS-TP node selects the traffic from one entity according to some predetermined criteria. Fault notification happens from the node identifying the fault to the root node and from the leaves to the root via an out of band path. In either case the root then selects the protection transport path for traffic transfer. More sophisticated survivability approaches such as partial tree protection and 1:n protection are for further study.

The IETF has no experience with P2MP PW survivability as yet, and therefore it is proposed that the P2MP PW survivability will initially rely on the LSP survivability. Further work is needed on this subject, particularly to if a requirement emerges to provide survivability for P2MP PWs in an MPLS-TP context.

7. Network Management

The network management architecture and requirements for MPLS-TP are specified in [RFC5951]. They derive from the generic specifications described in ITU-T G.7710/Y.1701 [G.7710] for transport technologies. They also incorporate the OAM requirements for MPLS Networks [RFC4377] and MPLS-TP Networks [RFC5860] and expand on those requirements to cover the modifications necessary for fault, configuration, performance, and security in a transport network.

[Editor's note: Decide what if anything needs to be said about P2MP-specific network management considerations.]

Section 3.14 of "Framework for MPLS in Transport Networks" [RFC5921] describe the aspects of network management in the P2P MPLS-TP case. This applies to the P2MP case. Packet Loss and Delay Measurement for MPLS Networks [RFC6374] already considers the P2MP case and it is not thought that any change is needed to the MPLS-TP profile of [RFC6374] [RFC6375].

8. Security Considerations

General security considerations for MPLS-TP are covered in [RFC5921]. Additional security considerations for point-to-multipoint LSPs are provided in [RFC4875]. This document introduces no new security considerations beyond those covered in those documents.

9. IANA Considerations

IANA considerations resulting from specific elements of MPLS-TP functionality are detailed in the documents specifying that functionality. This document introduces no additional IANA considerations in itself.

10. References

10.1. Normative References

- [RFC4872] Lang, J., Rekhter, Y., and D. Papadimitriou, "RSVP-TE Extensions in Support of End-to-End Generalized Multi-Protocol Label Switching (GMPLS) Recovery", RFC 4872, May 2007.
- [RFC4873] Berger, L., Bryskin, I., Papadimitriou, D., and A. Farrel, "GMPLS Segment Recovery", RFC 4873, May 2007.
- [RFC4875] Aggarwal, R., Papadimitriou, D., and S. Yasukawa, "Extensions to Resource Reservation Protocol - Traffic Engineering (RSVP-TE) for Point-to-Multipoint TE Label Switched Paths (LSPs)", RFC 4875, May 2007.
- [RFC5036] Andersson, L., Minei, I., and B. Thomas, "LDP Specification", RFC 5036, October 2007.
- [RFC5331] Aggarwal, R., Rekhter, Y., and E. Rosen, "MPLS Upstream Label Assignment and Context-Specific Label Space", RFC 5331, August 2008.

- [RFC5332] Eckert, T., Rosen, E., Aggarwal, R., and Y. Rekhter, "MPLS Multicast Encapsulations", RFC 5332, August 2008.
- [RFC5654] Niven-Jenkins, B., Brungard, D., Betts, M., Sprecher, N., and S. Ueno, "Requirements of an MPLS Transport Profile", RFC 5654, September 2009.
- [RFC5921] Bocci, M., Bryant, S., Frost, D., Levrau, L., and L. Berger, "A Framework for MPLS in Transport Networks", RFC 5921, July 2010.
- [RFC6374] Frost, D. and S. Bryant, "Packet Loss and Delay Measurement for MPLS Networks", RFC 6374, September 2011.
- [RFC6375] Frost, D. and S. Bryant, "A Packet Loss and Delay Measurement Profile for MPLS-Based Transport Networks", RFC 6375, September 2011.

10.2. Informative References

- [G.7710] "ITU-T Recommendation G.7710/Y.1701 (07/07), "Common equipment management function requirements"", 2005.
- [I-D.hmk-mpls-tp-p2mp-oam-framework] Koike, Y., Hamano, T., and M. Namiki, "A framework for Point-to-Multipoint MPLS-TP OAM in case that return paths don't exist", draft-hmk-mpls-tp-p2mp-oam-framework-00 (work in

progress), July 2012.

- [I-D.ietf-l2vpn-vpms-frmwk-requirements] Kamite, Y., JOUNAY, F., Niven-Jenkins, B., Brungard, D., and L. Jin, "Framework and Requirements for Virtual Private Multicast Service (VPMS)", draft-ietf-l2vpn-vpms-frmwk-requirements-04 (work in progress), July 2011.
- [I-D.ietf-pwe3-p2mp-pw] Sivabalan, S., Boutros, S., and L. Martini, "Signaling Root-Initiated Point-to-Multipoint Pseudowire using LDP", draft-ietf-pwe3-p2mp-pw-04 (work in progress), March 2012.
- [I-D.ietf-pwe3-p2mp-pw-requirements] Bocci, M., Heron, G., and Y. Kamite, "Requirements and Framework for Point-to-Multipoint Pseudowires over MPLS PSNs", draft-ietf-pwe3-p2mp-pw-requirements-05 (work in progress), September 2011.
- [I-D.ietf-pwe3-redundancy] Muley, P., Aissaoui, M., and M. Bocci, "Pseudowire Redundancy", draft-ietf-pwe3-redundancy-09 (work in progress), June 2012.
- [I-D.raggarwa-pwe3-p2mp-pw-encaps] Aggarwal, R. and F. JOUNAY, "Point-to-Multipoint Pseudo-Wire Encapsulation", draft-raggarwa-pwe3-p2mp-pw-encaps-01 (work in progress), March 2010.
- [RFC4377] Nadeau, T., Morrow, M., Swallow, G., Allan, D., and S. Matsushima, "Operations and Management (OAM) Requirements for Multi-

Protocol Label Switched
(MPLS) Networks", RFC 4377,
February 2006.

[RFC5860]

Vigoureux, M., Ward, D.,
and M. Betts, "Requirements
for Operations,
Administration, and
Maintenance (OAM) in MPLS
Transport Networks",
RFC 5860, May 2010.

[RFC5951]

Lam, K., Mansfield, S., and
E. Gray, "Network
Management Requirements for
MPLS-based Transport
Networks", RFC 5951,
September 2010.

[RFC6371]

Busi, I. and D. Allan,
"Operations,
Administration, and
Maintenance Framework for
MPLS-Based Transport
Networks", RFC 6371,
September 2011.

[RFC6372]

Sprecher, N. and A. Farrel,
"MPLS Transport Profile
(MPLS-TP) Survivability
Framework", RFC 6372,
September 2011.

[RFC6373]

Andersson, L., Berger, L.,
Fang, L., Bitar, N., and E.
Gray, "MPLS Transport
Profile (MPLS-TP) Control
Plane Framework", RFC 6373,
September 2011.

Authors' Addresses

Dan Frost (editor)
Cisco Systems

EMail: danfrost@cisco.com

Stewart Bryant (editor)
Cisco Systems

Phone:
Fax:
EMail: stbryant@cisco.com
URI:

Matthew Bocci (editor)
Alcatel-Lucent
Voyager Place, Shoppenhangers Road
Maidenhead, Berks SL6 2PJ
United Kingdom

EMail: matthew.bocci@alcatel-lucent.com

Lou Berger (editor)
LabN Consulting

Phone: +1-301-468-9228
EMail: lberger@labn.net

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: October 13, 2012

X. Fu
ZTE
V. Manral
Hewlett-Packard Corp.
D. McDysan
A. Malis
Verizon
S. Giacalone
Thomson Reuters
M. Betts
Q. Wang
ZTE
J. Drake
Juniper Networks
April 11, 2012

Loss and Delay Traffic Engineering Framework for MPLS
draft-fuxh-mpls-delay-loss-te-framework-05

Abstract

With more and more enterprises using cloud based services, the distances between the user and the applications are growing. A lot of the current applications are designed to work across LAN's and have various inherent assumptions. For multiple applications such as High Performance Computing and Electronic Financial markets, the response times are critical as is packet loss, while other applications require more throughput.

[RFC3031] describes the architecture of MPLS based networks. This draft extends the MPLS architecture to allow for latency, loss and jitter as properties. It describes requirements and control plane implication for latency and packet loss as a traffic engineering performance metric in today's network which is consisting of potentially multiple layers of packet transport network and optical transport network in order to make a accurate end-to-end latency and loss prediction before a path is established.

Note MPLS architecture for Multicast will be taken up in a future version of the draft.

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].

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 October 13, 2012.

Copyright Notice

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

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

Table of Contents

1. Introduction	4
2. Architecture requirements overview	4
2.1. Communicate Latency and Loss as TE Metric	4
2.2. Requirement for Composite Link	5
2.3. Requirement for Hierarchy LSP	5
2.4. Latency Accumulation and Verification	5
2.5. Restoration, Protection and Rerouting	6
3. End-to-End Latency	7
4. End-to-End Jitter	8
5. End-to-End Loss	8
6. Protocol Considerations	9
7. Control Plane Implication	10
7.1. Implications for Routing	10
7.2. Implications for Signaling	11
8. IANA Considerations	12
9. Security Considerations	13
10. Acknowledgements	13
11. References	13
11.1. Normative References	13
11.2. Informative References	13
Authors' Addresses	14

1. Introduction

In High Frequency trading for Electronic Financial markets, computers make decisions based on the Electronic Data received, without human intervention. These trades now account for a majority of the trading volumes and rely exclusively on ultra-low-latency direct market access.

Extremely low latency measurements for MPLS LSP tunnels are defined in [draft-ietf-mppls-loss-delay]. They allow a mechanism to measure and monitor performance metrics for packet loss, and one-way and two-way delay, as well as related metrics like delay variation and channel throughput.

The measurements are however effective only after the LSP is created and cannot be used by MPLS Path computation engine to define paths that have the latest latency. This draft defines the architecture used, so that end-to-end tunnels can be set up based on latency, loss or jitter characteristics.

End-to-end service optimization based on latency and packet loss is a key requirement for service provider. This type of function will be adopted by their "premium" service customers. They would like to pay for this "premium" service. Latency and loss on a route level will help carriers' customers to make his provider selection decision.

2. Architecture requirements overview

2.1. Communicate Latency and Loss as TE Metric

The solution MUST provide a means to communicate latency, latency variation and packet loss of links and nodes as a traffic engineering performance metric into IGP.

Latency, latency variation and packet loss may be unstable, for example, if queueing latency were included, then IGP could become unstable. The solution MUST provide a means to control latency and loss IGP message advertisement rate and avoid instability when the latency, latency variation and packet loss value changes frequently.

In the case where it is known that either the changes are too frequent or there is a backup which is preferred, the solution shall put the node or the link in unusable state for services requiring a particular service capability. This unusable state is on a capability basis and not a global basis. The condition to get into the state is locally configured and all routers in a domain should have this criteria synchronized.

Path computation entity MUST have the capability to compute one end-to-end path with latency and packet loss constraint. For example, it has the capability to compute a route with X amount of bandwidth with less than Y ms of latency and less than Z% packet loss limit based on the latency and packet loss traffic engineering database. It MUST also support the path computation with routing constraints combination with pre-defined priorities, e.g., SRLG diversity, latency, loss, jitter and cost. If the performance of link exceeds its configured maximum threshold, path computation entity may not select this kind of link although end-to-end performance is still met.

2.2. Requirement for Composite Link

One end-to-end LSP may traverses some Composite Links [CL-REQ]. Even if the transport technology (e.g., OTN) component links are identical, the latency and packet loss characteristics of the component links may differ due to factors such as fiber distance and/or fiber characteristics.

The solution MUST provide a means to indicate that a traffic flow should select a component link with minimum latency and/or packet loss, maximum acceptable latency and/or packet loss value and maximum acceptable delay variation value as specified by protocol. The endpoints of Composite Link will take these parameters into account for component link selection or creation. Details of how transient response is taken is specified in Section 4.1 [CL-REQ]. The exact details for component links will be taken up separately and are not part of this document.

2.3. Requirement for Hierarchy LSP

Heirarchical LSP's may traverse server layer LSP's. For such LSP's there may be some latency and packet loss constraint requirement for the segment in server layer.

The solution MUST provide a means to indicate FA selection or FA-LSP creation with minimum latency and/or packet loss, maximum acceptable latency and/or packet loss value and maximum acceptable delay variation value. The boundary nodes of FA-LSP will take these parameters into account for FA selection or FA-LSP creation.

2.4. Latency Accumulation and Verification

The solution SHOULD provide a means to accumulate (e.g., sum) latency information of links and nodes along that an LSP traverses, (e.g., Inter-AS, Inter-Area or Multi-Layer) so that the source node can validate if the desired maximum latency constraint can be satisfied

for a packet traversing the LSP. [Y.1541] provides details of how the latency value is accumulated.

Both One-way and Round-trip latency collection along the LSP by signaling protocol and latency verification at the end of LSP should be supported.

The accumulation of the delay is "simple" for the static component i.e. its a linear addition, the dynamic/network loading component is more interesting and would involve some estimate of the "worst case". However, method of deriving this worst case appears to be more in the scope of Network Operator policy than standards i.e. the operator needs to decide, based on the SLAs offered, the required confidence level.

2.5. Restoration, Protection and Rerouting

Some customers may insist on having the ability to re-route if the latency and loss SLA is not being met. If a "provisioned" end-to-end LSP latency and/or loss could not meet the latency and loss agreement between operator and his user, the solution SHOULD support pre-defined or dynamic re-routing (e.g., make-before-break) to handle this case based on the local policy. In revertive behaviour is supported, the original LSP must not be released and is monitored by control plane. When the end-to-end performance is repaired, the service is restored to the original LSP.

The solution SHOULD support to move an end-to-end LSP away from any link whose performance violates the configured threshold.

End-to-end measurements of the LSP also need to be performed in addition to the link-by-link measurements. A threshold violation of the End-to-End criteria as measured by the head end node should cause rerouting of the LSP.

The anomalous path can be switch to protection path or rerouted to new path because of end-to-end performance couldn't meet any more.

If a "provisioned" end-to-end LSP latency and/or loss performance is improved (i.e., beyond a configurable minimum value), the solution SHOULD support the re-routing to optimize latency and/or loss end-to-end cost.

The latency performance of pre-defined protection or dynamic re-routing LSP MUST meet the latency SLA parameter. The difference of latency, jitter or loss value between primary and protection/restoration path SHOULD be zero. [MPLS-TP-USE-CASE] defines a Relative Delay Time which is the difference of the Absolute Delay

Time between using working and protect path. When the relative network latency is increased or decreased, the customer would complain. From network operational point of view, they want to minimize the number of customers complains. The scope of this draft is much broader than MPLS-TP and there is a need for a framework to identify all of these related requirements.

Due to some flapping conditions the latency and loss of an LSP may change, this may cause the LSP to be frequently switched to a new path. In order to avoid churn, the solution SHOULD specify the switchover of the LSP according to maximum acceptable change rate.

3. End-to-End Latency

Procedures to measure latency and loss has been provided in ITU-T [Y.1731], [G.709] and [ietf-mpls-loss-delay]. The control plane can be independent of the mechanism used and different mechanisms can be used for measurement based on different standards.

Latency on a path has two sources: Node latency which is caused by the node as a result of process time in each node and: Link latency as a result of packet/frame transit time between two neighbouring nodes or a FA-LSP/ Composite Link [CL-REQ].

Latency or one-way delay is the time it takes for a packet within a stream going from measurement point 1 to measurement point 2, as defined in [Y.1540].

The architecture uses assumption that the sum of the latencies of the individual components approximately adds up to the average latency of an LSP. Though using the sum may not be perfect, it however gives a good approximation that can be used for Traffic Engineering (TE) purposes.

The total measured latency of an LSP consists of the sum of the latency of the LSP hop, as well as the average latency of switching on a device, which may vary based on queuing and buffering.

Hop latency can be measured by getting the latency measurement between the egress of one MPLS LSR to the ingress of the nexthop LSR. This value may be constant for most part, unless there is protection switching, or other similar changes at a lower layer.

The switching latency on a device, can be measured internally, and multiple mechanisms and data structures to do the same have been defined. [Add references to papers by Verghese, Kompella, Duffield].

We also looked at other measurement granularities before deciding on an interface based measurement. An approximation of the Flow based measurement is the per DSCP value, measurement from the ingress of one port to the egress of every other port in the device.

Another approximation that can be used is per interface DSCP based measurement, which can be an aggregate of the average measurements per interface. The average can itself be calculated in ways, so as to provide closer approximation.

For the purpose of this draft it is assumed that the node latency is a small factor of the total latency in the networks where this solution is deployed. The node latency is hence ignored for the benefit of simplicity in this solution.

The average link delay over a configurable interval should be reported by data plane in micro-seconds.

4. End-to-End Jitter

Jitter or Packet Delay Variation of a packet within a stream of packets is defined for a selected pair of packets in the stream going from measurement point 1 to measurement point 2.

This architecture uses the assumptions of [Y.1540] to calculate the accumulated jitter from the individual components approximately. Though using this may not be perfect, it however gives a good approximation that can be used for Traffic Engineering (TE) purposes.

The buffering and queuing within a device will lead to the jitter. Just like latency measurements, jitter measurements can be approximated as either per DSCP per port pair (Ingress and Egress) or as per DSCP per egress port, however such measurements have been left out for the sake of simplicity of the solution.

For the purpose of this draft it is assumed that the node latency is a small factor of the total latency in the networks where this solution is deployed. The node latency is hence ignored for the benefit of simplicity.

The jitter is measured in micro-seconds.

5. End-to-End Loss

Loss or Packet Drop probability of a packet within a stream of packets is defined as the number of packets dropped within a given

interval.

This architecture uses the assumptions of [Y.1540] to calculate the accumulated loss from the individual components approximately. Though using the accumulated metrics may not be perfect, it however gives a good approximation that can be used for Traffic Engineering (TE) purposes.

The buffering and queuing mechanisms within a device will decide which packet is to be dropped. Just like latency and jitter measurements, the loss can best be approximated as either per DSCP per port pair (Ingress and Egress) or as per DSCP per egress port. However such mechanisms are not used in this solution to keep the solution simple.

The loss is measured in terms of the number of packets per million packets.

6. Protocol Considerations

The protocol metrics above can be sent in IGP protocol packets as defined in [OSPF-TE-EXPRESS-PATH] and [ISIS-TE-EXPRESS-PATH]. They can then be used by Source Node or the Path Computation engine to decide paths with the desired path properties. [EXPRESS-PATH] describes how to use these traffic engineering metrics to compute explicit paths at path computation entity.

As Link-state IGP information is flooded throughout an area, frequent changes can cause a lot of control traffic. To prevent such flooding, data should only be flooded when it crosses a certain configured maximum.

A separate measurement should be done for an LSP when it is UP. Also LSP's path should only be recalculated when the end-to-end metrics changes in a way it becomes more than desired.

Delay, jitter or loss is part of service/QoS description/characterization. RSVP-TE extension is defined in [DELAY-LOSS-RSVP-TE].

This document is a framework tracking the various solution approaches and placing them in context. This document additionally provides a framework for the control of MPLS networks based on delay and loss TE.

7. Control Plane Implication

7.1. Implications for Routing

The latency and packet loss performance metric **MUST** be advertised into path computation entity by IGP (OSPF-TE, OSPFv3-TE or IS-IS-TE) to perform route computation and network planning based on latency and packet loss SLA target.

Latency, latency variation and packet loss value **MUST** be reported as a average value which is calculated by data plane measurements.

Latency and packet loss characteristics of these links and nodes may change dynamically. In order to control IGP messaging and avoid being unstable when the latency, latency variation and packet loss value changes, a threshold and a limit on rate of change **MUST** be configured in the IGP control plane.

Latency and packet loss values changes need to be updated and flooded in the IGP control messages only when there is significant changes in the value. When the head end-node determines the IGP update affects the LSP for which it is ingress, it recalculates the LSP.

A target value **MUST** be configured to control plane for each link. If the link performance improves beyond a configurable target value, it must be re-advertised. The receiving node determines whether a "provisioned" end-to-end LSP latency and/or loss performance is improved.

It is sometimes important for paths that desire low latency to avoid nodes that have a significant contribution to latency. Control plane should report two components of the delay, "static" and "dynamic". The dynamic component is always caused by traffic loading and queuing. The "dynamic" portion **SHOULD** be reported as an approximate value. The static component should be a fixed latency through the node without any queuing. Link latency attribute should also take into account the latency of node, i.e., the latency between the incoming port and the outgoing port of a network element. Half of the fixed node latency can be added to each link.

When the Composite Links [CL-REQ] is advertised into IGP, there are following considerations.

- o One option is that the latency and packet loss of composite link may be the range (e.g., at least minimum and maximum) latency value of all component links. It may also be the maximum or average latency value of all component links. In both cases, only partial information is transmitted in the IGP. So the path

computation entity has insufficient information to determine whether a particular path can support its latency and packet loss requirements. This leads to signaling crankback.

- o Another option is that latency and packet loss of each component link within one Composite Link could be advertised but having only one IGP adjacency.

One end-to-end LSP (e.g., in IP/MPLS or MPLS-TP network) may traverse a FA-LSP of server layer (e.g., OTN rings). The boundary nodes of the FA-LSP SHOULD be aware of the latency and packet loss information of this FA-LSP.

If the FA-LSP is able to form a routing adjacency and/or as a TE link in the client network, the total latency and packet loss value of the FA-LSP can be as an input to a transformation that results in a FA traffic engineering metric and advertised into the client layer routing instances. Note that this metric will include the latency and packet loss of the links and nodes that the trail traverses.

If total latency and packet loss information of the FA-LSP changes (e.g., due to a maintenance action or failure in OTN rings), the boundary node of the FA-LSP will receive the TE link information advertisement including the latency and packet value which is already changed and if it is over than the threshold and a limit on rate of change, then it will compute the total latency and packet value of the FA-LSP again. If the total latency and packet loss value of FA-LSP changes, the client layer MUST also be notified about the latest value of FA. The client layer can then decide if it will accept the increased latency and packet loss or request a new path that meets the latency and packet loss requirement.

7.2. Implications for Signaling

In order to assign the LSP to one of component links with different latency and loss characteristics, RSVP-TE message needs to carry a indication of request minimum latency and/or packet loss, maximum acceptable latency and/or packet loss value and maximum acceptable delay variation value for the component link selection or creation. The composite link will take these parameters into account when assigning traffic of LSP to a component link.

One end-to-end LSP (e.g., in IP/MPLS or MPLS-TP network) may traverse a FA-LSP of server layer (e.g., OTN rings). There will be some latency and packet loss constraint requirement for the segment route in server layer. So RSVP-TE message needs to carry a indication of request minimum latency and/or packet loss, maximum acceptable latency and/or packet loss value and maximum acceptable delay

variation value. The boundary nodes of FA-LSP will take these parameters into account for FA selection or FA-LSP creation.

RSVP-TE needs to be extended to accumulate (e.g., sum) latency information of links and nodes along one LSP across multi-domain (e.g., Inter-AS, Inter-Area or Multi-Layer) so that a latency verification can be made at end points. One-way and round-trip latency collection along the LSP by signaling protocol can be supported. So the end points of this LSP can verify whether the total amount of latency could meet the latency agreement between operator and his user. When RSVP-TE signaling is used, the source can determine if the latency requirement is met much more rapidly than performing the actual end-to-end latency measurement.

Restoration, protection and equipment variations can impact "provisioned" latency and packet loss (e.g., latency and packet loss increase). For example, restoration/provisioning action in transport network that increases latency seen by packet network observable by customers, possibly violating SLAs. The change of one end-to-end LSP latency and packet loss performance MUST be known by source and/or sink node. So it can inform the higher layer network of a latency and packet loss change. The latency or packet loss change of links and nodes will affect one end-to-end LSPs total amount of latency or packet loss. Applications can fail beyond an application-specific threshold. Some remedy mechanism could be used.

Pre-defined protection or dynamic re-routing could be triggered to handle this case. In the case of predefined protection, large amounts of redundant capacity may have a significant negative impact on the overall network cost. Service provider may have many layers of pre-defined restoration for this transfer, but they have to duplicate restoration resources at significant cost. Solution should provide some mechanisms to avoid the duplicate restoration and reduce the network cost. Dynamic re-routing also has to face the risk of resource limitation. So the choice of mechanism MUST be based on SLA or policy. In the case where the latency SLA can not be met after a re-route is attempted, control plane should report an alarm to management plane. It could also try restoration for several times which could be configured.

8. IANA Considerations

No new IANA consideration are raised by this document.

9. Security Considerations

This document raises no new security issues.

10. Acknowledgements

TBD.

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.
- [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.
- [RFC3477] Kompella, K. and Y. Rekhter, "Signalling Unnumbered Links in Resource ReSerVation Protocol - Traffic Engineering (RSVP-TE)", RFC 3477, January 2003.
- [RFC3630] Katz, D., Kompella, K., and D. Yeung, "Traffic Engineering (TE) Extensions to OSPF Version 2", RFC 3630, September 2003.
- [RFC4203] Kompella, K. and Y. Rekhter, "OSPF Extensions in Support of Generalized Multi-Protocol Label Switching (GMPLS)", RFC 4203, October 2005.

11.2. Informative References

- [CL-REQ] C. Villamizar, "Requirements for MPLS Over a Composite Link", draft-ietf-rtgwg-cl-requirement-04 .
- [DELAY-LOSS-RSVP-TE] X. Fu, "RSVP-TE extensions for Delay and Loss Traffic Engineering", draft-fuxh-mpls-delay-loss-rsvp-te-ext-01 .
- [EXPRESS-PATH] A. Atlas, "Performance-based Path Selection for Explicitly

Routed LSPs", draft-atlas-mpls-te-express-path-00 .

[G.709] ITU-T Recommendation G.709, "Interfaces for the Optical Transport Network (OTN)", December 2009.

[ISIS-TE-EXPRESS-PATH]
S. Previdi, "IS-IS Traffic Engineering (TE) Metric Extensions", draft-ietf-ospf-te-metric-extensions-00 .

[MPLS-TP-USE-CASE]
L. Fang, "MPLS-TP Applicability; Use Cases and Design", draft-ietf-mpls-tp-use-cases-and-design-01 .

[OSPF-TE-EXPRESS-PATH]
S. Giacalone, "OSPF Traffic Engineering (TE) Metric Extensions", draft-ietf-ospf-te-metric-extensions-00 .

[Y.1731] ITU-T Recommendation Y.1731, "OAM functions and mechanisms for Ethernet based networks", Feb 2008.

[ietf-mpls-loss-delay]
D. Frost, "Packet Loss and Delay Measurement for MPLS Networks", draft-ietf-mpls-loss-delay-03 .

Authors' Addresses

Xihua Fu
ZTE

Email: fu.xihua@zte.com.cn

Vishwas Manral
Hewlett-Packard Corp.
191111 Pruneridge Ave.
Cupertino, CA 95014
US

Phone: 408-447-1497
Email: vishwas.manral@hp.com
URI:

Dave McDysan
Verizon

Email: dave.mcdysan@verizon.com

Andrew Malis
Verizon

Email: andrew.g.malis@verizon.com

Spencer Giacalone
Thomson Reuters
195 Broadway
New York, NY 10007
US

Phone: 646-822-3000
Email: spencer.giacalone@thomsonreuters.com
URI:

Malcolm Betts
ZTE

Email: malcolm.betts@zte.com.cn

Qilei Wang
ZTE

Email: wang.qilei@zte.com.cn

John Drake
Juniper Networks

Email: jdrake@juniper.net

MPLS Working Group
Internet Draft

Y.Koike, Ed.
T.Hamano
M.Namiki
NTT

Intended status: Informational

Expires: January 8, 2013

July 9, 2012

A framework for Point-to-Multipoint MPLS-TP OAM in case that return
paths don't exist
draft-hmk-mpls-tp-p2mp-oam-framework-00.txt

Status of this Memo

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

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

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

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

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

This Internet-Draft will expire on January 8, 2013.

Copyright Notice

Copyright (c) 2011 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents

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

Abstract

The MPLS transport profile (MPLS-TP) is being standardized to enable carrier-grade packet transport.

This document provides texts proposal which should be discussed and included in draft-fbb-mpls-tp-p2mp-framework, particularly focusing on p2mp OAM framework in case that return paths don't exist. Other requirements of p2mp transport path such as protection will also be discussed.

Note: This I-D was made based on the result of discussion in ITU-T SG15 which is described in a Liaison Statement: Request advance work on the p2mp framework in MPLS-TP
(<https://datatracker.ietf.org/liaison/1163/>)

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

Table of Contents

1. Introduction	3
2. Conventions used in this document.....	4
2.1. Terminology	4
2.2. Definitions	4
3. P2MP OAM	4
3.1. OAM functions for proactive monitoring	5
3.1.1. Continuity Check and Connectivity Verification.....	5
3.1.2. Remote Defect Indication.....	6

3.1.3. Alarm Reporting.....	6
3.1.4. Lock Reporting.....	7
3.1.5. Packet Loss Measurement.....	7
3.1.6. Packet Delay Measurement.....	7
3.1.7. Client Failure Indication	7
3.2. OAM functions for on-demand monitoring	7
3.2.1. Connectivity verification	7
3.2.2. Packet loss measurement.....	8
3.2.3. Diagnostic tests.....	9
3.2.4. Route Tracing.....	9
3.2.5. Packet delay measurement.....	9
3.3. OAM functions for administration control.....	9
3.3.1. Lock Instruct.....	9
4. Security Considerations.....	9
5. IANA Considerations	9
6. References	9
6.1. Normative References.....	9
6.2. Informative References.....	10
7. Acknowledgments	10

1. Introduction

The demand for P2MP traffic is expected to increase due to the increase in new services such as IP-TV and video distribution services. Moreover considering the global trend to improve energy efficiency, a P2MP transport function in MPLS-TP could be one of the solutions to achieve this goal from the perspective of efficient use of network resources.

RFC5654[1] defines the following requirements which are specific to P2MP.

- Traffic-engineered point-to-multipoint (P2MP) transport paths.(item 6)
- Unidirectional point-to-multipoint transport paths (item 8)
- Being capable of using P2MP server (sub)layer capabilities when supporting P2MP MPLS-TP transport paths(item 40)
- The MPLS-TP control plane MUST support establishing all the connectivity patterns defined for the MPLS-TP data plane (i.e. unidirectional P2MP) including configuration of protection functions and any associated maintenance functions.(item 50)
- Unidirectional 1+1 protection for P2MP connectivity (item 65 C)
- Unidirectional 1:n protection for P2MP connectivity(item 67 B)
- MPLS-TP recovery in a ring MUST protect unidirectional P2MP transport paths.(item 95)

RFC5860[2] defines MPLS-TP OAM requirements including those for unidirectional P2MP transport paths. In case of unidirectional P2MP transport path, two cases are assumed as per section 3.3 of RFC6371[3]. One is when an "out-of-band" return path exists and it is used and the other is when any return path does not exist or is not used. Missing OAM requirements which are necessary in P2MP transport networks are those in the latter case.

In I-D[4], Operations, Administration and Maintenance (OAM) is planned to be specified in clause 4. According to the editor's note, this section will contain a summary of point-to-multipoint OAM as described in RFC6371[3] that defines the overall OAM architecture for MPLS-TP.

However, considering the missing OAM requirements in case that a return path doesn't exist, the most appropriate place where they could be added is I-D[4]. Therefore, this draft intends to provide texts which should be included in OAM section and network management section of the I-D[4].

2. Conventions used in this document

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC-2119 [1].

2.1. Terminology

LSP Label Switched Path

2.2. Definitions

None

3. P2MP OAM

Note: It is proposed that this section be incorporated in section 4 of I-D[4].

Unidirectional P2MP is supported in MPLS-TP. This means that "in-band" return path is out of scope. In this section, only two cases,

with out-band return path and without return path, are considered and requirements should be independently specified, if necessary.

P2MP considerations are described in section 3.7 of RFC6371. The RFC has already described some requirements with out-band return path(s). On the other hand, even if there is no return path, parts of OAM requirements in RFC5860 could be met by supporting management interface through which EMS/NMS can retrieve the received OAM packets.

Note: In the following sections, basically additional requirements are described function-by-function, which haven't been covered or clarified in RFC5860[2] and RFC6371[3] have particularly focused on the case that return paths don't exist.

3.1. OAM functions for proactive monitoring

3.1.1. Continuity Check and Connectivity Verification

Continuity Check function enable one or more leaf MEPs on unidirectional P2MP transport path to monitor the continuity of OAM packets from root MEP and detect one or more loss of continuity(LOC) defect between the root MEP and the leaf MEPs. Connectivity Verification function enables one or more leaf MEPs on P2MP transport path to monitor the connectivity of OAM packets from a specific root MEP and detect an unexpected connectivity defect between two MEGs(two P2MP transport paths)

Continuity Check and Connectivity Verification MUST be supported in case that a return path in a unidirectional P2MP transport path doesn't exist. This requirement is already included in section 2.2.3 of RFC5860[2].

As described in RFC6371[3], CC-V OAM packets are used for P2MP transport path. Defect detection mechanisms in P2MP transport paths are the same as those of P2MP transport path specified in section 5.1 of RFC6371. That is, loss of continuity defect, mis-connectivity defect, period mis-configuration defect and unexpected encapsulation defect. Entry criteria and exist criteria are also the same as those of P2MP transport path in RFC6371[3]. Moreover, consequent actions of unidirectional P2MP transport path are also covered in section 5.1.2 of the RFC[3]

Regarding configuration consideration, following additional requirements on unidirectional P2MP transport path in case that the return paths don't exist.

1. EMS/NMS should provide a tool to manually configure consistent values on each piece of configuration information (MEG-ID, MEP-ID, list of the other MEPs in the MEG, PHB for E-LSPs, transmission rate) to a root-MEP and all the related leaf-MEPs in a MEG of a P2MP transport path.
2. Mis-matches of configuration information (MEG-ID, MEP-ID, PHB for E-LSPs, transmission rate) between a root MEP and any leaf-MEP at which proactive monitoring is enabled, should be detected as a configuration mis-match alarm by parsing received CC-VOAM packets.
3. Mis-matches of configuration information (MEG-ID, MEP-ID, list of the other MEPs in the MEG, PHB for E-LSPs, transmission rate) between a leaf MEP and any other leaf-MEP, at which proactive monitoring are enabled, may be detected through configuration management process of EMS/NMS as a configuration mis-match alarm without receiving OAM packets from a source MEP.
4. Configuration information mis-match alarms described in 4 and 5 may be supported in case that a proactive monitoring is not enabled in order to check those mis-matches before monitoring functions are enabled.
5. Enabling or disabling configuration mis-match alarms must be able to be configured at each leaf-MEP independently.

3.1.2. Remote Defect Indication

This OAM function is not available on P2MP transport path in case that return paths don't exist, because this function is implemented only on the return path.

3.1.3. Alarm Reporting

Alarm Reporting functions MUST be supported in case that a return path in a unidirectional P2MP transport paths don't exist. This is already included in section 2.2.8 of RFC5860[2].

6. EMS/NMS should provide a tool to manually configure consistent values on "hold-off intervals prior to asserting an alarm to the management system" and AIS transmission period to all the leaf-MEPs in a MEG of a P2MP transport path.
7. Mis-matches of configuration information (hold-off interval and AIS transmission period) between a root MEP and any leaf-MEP at which alarm reporting is enabled, should be detected as a configuration mis-match alarm by parsing received AIS OAM packets.

8. Mis-matches of configuration information (hold-off interval and AIS transmission period) between a leaf MEP and any other leaf-MEP, at which alarm reporting is enabled, may be detected through configuration management process of EMS/NMS as a configuration mis-match alarm without receiving OAM packets from a source MEP.
9. Configuration information mis-match alarms described in 4 and 5 may be supported in case that a alarm reporting is not enabled in order to check those mis-matches before monitoring functions are enabled.
10. Enabling or disabling configuration information mis-match alarms must be able to be configured at each leaf-MEP independently.

3.1.4. Lock Reporting

FFS

3.1.5. Packet Loss Measurement

FFS

3.1.6. Packet Delay Measurement

FFS

3.1.7. Client Failure Indication

FFS

3.2. OAM functions for on-demand monitoring

3.2.1. Connectivity verification

Connectivity Verification function enables one or more leaf MEPs on P2MP transport path to monitor the connectivity of OAM packets from a specific root MEP and detect an unexpected connectivity defect between two MEGs (two P2MP transport paths)

11. Connectivity verification functions MUST be supported in case that return paths in a unidirectional P2MP transport path don't exist.

As described in RFC6371[3], CC-V OAM packets are used for P2MP transport path. Defect detection mechanisms in P2MP transport paths are the same as those of P2MP transport path specified in section 5.1 of RFC6371. That is, loss of continuity defect, mis-connectivity defect, period mis-configuration defect and unexpected encapsulation defect. Entry criteria and exist criteria are also the same as those of P2MP transport path in RFC6371[3]. Moreover, consequent actions of unidirectional P2MP transport path are also covered in section 5.1.2 of the RFC[3]

Regarding configuration consideration, following additional requirements on unidirectional P2MP transport path in case that return path doesn't exist.

- 12.EMS/NMS should provide a tool to manually configure consistent values on each piece of configuration information (MEG-ID, MEP-ID, list of the other MEPs in the MEG, PHB for E-LSPs, transmission rate) to a root-MEP and all the related leaf-MEPs in a MEG of a P2MP transport path.
- 13.Mis-matches of configuration information (MEG-ID, MEP-ID, PHB for E-LSPs, transmission rate) between a root MEP and any leaf-MEP at which proactive monitoring is enabled, should be detected as a configuration mis-match alarm by parsing received CC-VOAM packets.
- 14.Mis-matches of configuration information (MEG-ID, MEP-ID, list of the other MEPs in the MEG, PHB for E-LSPs, transmission rate) between a leaf MEP and any other leaf-MEP, at which proactive monitoring are enabled, may be detected through configuration management process of EMS/NMS as a configuration mis-match alarm without receiving OAM packets from a source MEP.
- 15.Configuration information mis-match alarms described in 4 and 5 may be supported in case that a proactive monitoring is not enabled in order to check those mis-matches before monitoring functions are enabled.
- 16.Enabling or disabling configuration mis-match alarms must be able to be configured at each leaf-MEP independently.

3.2.2. Packet loss measurement

FFS

3.2.3. Diagnostic tests

17. Diagnostic test functions MUST be supported in case that a return path in a unidirectional P2MP transport path doesn't exist.

Other requirements are ffs.

3.2.4. Route Tracing

18. Route tracing function MUST be supported in case that a return path in a unidirectional P2MP transport path doesn't exist.

Other requirements are ffs.

3.2.5. Packet delay measurement

FFS

3.3. OAM functions for administration control

3.3.1. Lock Instruct

FFS.

4. Security Considerations

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

5. IANA Considerations

There are no IANA actions required by this draft.

6. References

6.1. Normative References

- [1] Niven-Jenkins, B., et al, "Requirements of an MPLS Transport Profile", RFC5654, September 2009
- [2] Vigoureux, M., Betts, M., Ward, D., "Requirements for OAM in MPLS Transport Networks", RFC5860, May 2010

- [3] Busi, I., Dave, A. , "Operations, Administration and Maintenance Framework for MPLS-based Transport Networks ", RFC6371, September 2011
- [4] Frost, Dan., et all, "A Framework for Point-to-Multipoint MPLS in Transport Networks", draft-fbb-mpls-tp-p2mp-framework-04, June 2012

6.2. Informative References

None

7. Acknowledgments

The author would like to thank all members (including MPLS-TP steering committee, the Joint Working Team, the MPLS-TP Ad Hoc Group in ITU-T) involved in the definition and specification of MPLS Transport Profile.

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

Authors' Addresses

Takafumi Hamano
NTT
hamano.takafumi@lab.ntt.co.jp

Masatoshi Namiki
NTT
namiki.masatoshi@lab.ntt.co.jp

Yoshinori Koike
NTT
Email: koike.yoshinori@lab.ntt.co.jp

MPLS Working Group
Internet Draft
Updates: 5036 (if approved)
Intended status: Standards Track
Expires: December 8, 2012

Rajiv Asati
Cisco

Vishwas Manral
Hewlett-Packard, Inc.

Rajiv Papneja
Huawei

Carlos Pignataro
Cisco

June 8, 2012

Updates to LDP for IPv6
draft-ietf-mpls-ldp-ipv6-07

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 8, 2012.

Copyright Notice

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

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this

document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Abstract

The Label Distribution Protocol (LDP) specification defines procedures to exchange label bindings over either IPv4, IPv6 or both networks. This document corrects and clarifies the LDP behavior when IPv6 network is used (with or without IPv4). This document updates RFC 5036.

Table of Contents

1. Introduction.....	3
1.1. Scope.....	4
1.1.1. Topology Scenarios.....	4
1.1.2. LDP TTL Security.....	5
2. Specification Language.....	5
3. LSP Mapping.....	6
4. LDP Identifiers.....	6
5. Peer Discovery.....	7
5.1. Basic Discovery Mechanism.....	7
5.2. Extended Discovery Mechanism.....	8
6. LDP Session Establishment and Maintenance.....	8
6.1. Transport connection establishment.....	9
6.2. Maintaining Hello Adjacencies.....	10
6.3. Maintaining LDP Sessions.....	11
7. Label Distribution.....	11
8. LDP Identifiers and Next Hop Addresses.....	12
9. LDP TTL Security.....	13
10. IANA Considerations.....	14
11. Security Considerations.....	14
12. Acknowledgments.....	14
13. Additional Contributors.....	15
14. References.....	16
14.1. Normative References.....	16
14.2. Informative References.....	16
Author's Addresses.....	17

1. Introduction

The LDP [RFC5036] specification defines procedures and messages for exchanging FEC-label bindings over either IPv4 or IPv6 or both (e.g. dual-stack) networks.

However, RFC5036 specification has the following deficiencies in regards to IPv6 usage:

- 1) LSP Mapping: No rule defined for mapping a particular packet to a particular LSP that has an Address Prefix FEC element containing IPv6 address of the egress router
- 2) LDP Identifier: No details specific to IPv6 usage
- 3) LDP Discovery: No details for using a particular IPv6 destination (multicast) address or the source address (with or without IPv4 co-existence)
- 4) LDP Session establishment: No rule for handling both IPv4 and IPv6 transport address optional objects in a Hello message, and subsequently two IPv4 and IPv6 transport connections
- 5) LDP Label Distribution: No rule for advertising IPv4 or/and IPv6 FEC-label bindings over an LDP session, and denying the co-existence of IPv4 and IPv6 FEC Elements in the same FEC TLV
- 6) Next Hop Address & LDP Identifier: No rule for accommodating the usage of duplicate link-local IPv6 addresses
- 7) LDP TTL Security: No rule for built-in Generalized TTL Security Mechanism (GTSM) in LDP

This document addresses the above deficiencies by specifying the desired behavior/rules/details for using LDP in IPv6 enabled networks. It also clarifies the scope (section 1.1).

Note that this document updates RFC5036.

1.1. Scope

1.1.1. Topology Scenarios

The following scenarios in which the LSRs may be inter-connected via one or more dual-stack interfaces (figure 1), or two or more single-stack interfaces (figure 2 and figure 3) are addressed by this document:

R1-----R2
IPv4+IPv6

Figure 1 LSRs connected via a Dual-stack Interface

IPv4
R1=====R2
IPv6

Figure 2 LSRs connected via two single-stack Interfaces

R1-----R2-----R3
IPv4 IPv6

Figure 3 LSRs connected via a single-stack Interface

Note that the topology scenario illustrated in figure 1 also covers the case of a single-stack interface (IPv4, say) being converted to a dual-stacked interface by enabling IPv6 as well as IPv6 LDP, even though the IPv4 LDP session may already be established between the LSRs.

Note that the topology scenario illustrated in figure 2 also covers the case of two routers getting connected via an additional single-stack interface (IPv6, say), even though the IPv4 LDP session may already be established between the LSRs over the existing interface.

1.1.2. LDP TTL Security

LDP TTL Security mechanism specified by this document applies only to single-hop LDP peering sessions, but not to multi-hop LDP peering sessions, in line with Section 5.5 of [RFC5082] that describes Generalized TTL Security Mechanism (GTSM).

As a consequence, any LDP feature that relies on multi-hop LDP peering session would not work with GTSM and will warrant (statically or dynamically) disabling GTSM. Please see section 8.

2. Specification 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].

Abbreviations:

LDP	- Label Distribution Protocol
LDPv4	- LDP for enabling IPv4 MPLS forwarding
LDPv6	- LDP for enabling IPv6 MPLS forwarding
LDPoIPv4	- LDP over IPv4 transport session
LDPoIPv6	- LDP over IPv6 transport session
FEC	- Forwarding Equivalence Class
TLV	- Type Length Value
LSR	- Label Switch Router
LSP	- Label Switched Path
LSPv4	- IPv4-signaled Label Switched Path [RFC4798]
LSPv6	- IPv6-signaled Label Switched Path [RFC4798]

3. LSP Mapping

Section 2.1 of [RFC5036] specifies the procedure for mapping a particular packet to a particular LSP using three rules. Quoting the 3rd rule from RFC5036:

"If it is known that a packet must traverse a particular egress router, and there is an LSP that has an Address Prefix FEC element that is a /32 address of that router, then the packet is mapped to that LSP."

Suffice to say, this rule is correct for IPv4, but not for IPv6, since an IPv6 router may not have any /32 address.

This document proposes to modify this rule by also including a /128 address (for IPv6). In fact, it should be reasonable to just say IPv4 or IPv6 address instead of /32 or /128 addresses as shown below in the updated rule:

"If it is known that a packet must traverse a particular egress router, and there is an LSP that has an Address Prefix FEC element that is an IPv4 or IPv6 address of that router, then the packet is mapped to that LSP."

Additionally, it is desirable that a packet is forwarded to an LSP of an egress router, only if LSP's address-family (e.g. LSPv4 or LSPv6) matches with that of the LDP hello adjacency on the next-hop interface.

4. LDP Identifiers

Section 2.2.2 of [RFC5036] specifies formulating at least one LDP Identifier, however, it doesn't provide any consideration in case of IPv6 (with or without dual-stacking). Additionally, section 2.5.2 of [RFC5036] implicitly prohibits using the same label space for both IPv4 and IPv6 FEC-label bindings.

The first four octets of the LDP identifier, the 32-bit LSR Id (e.g. (i.e. LDP Router Id), identify the LSR and is a globally unique value within the MPLS network. This is regardless of the address family used for the LDP session. Hence, this document preserves the usage of 32-bit (unsigned non-zero integer) LSR Id on an IPv6 only

LSR (note that BGP has also mandated using 32-bit BGP Router ID on an IPv6 only Router [RFC6286]).

Please note that 32-bit LSR Id value would not map to any IPv4-address in an IPv6 only LSR (i.e., single stack), nor would there be an expectation of it being DNS-resolvable. In IPv4 deployments, the LSR Id is typically derived from an IPv4 address, generally assigned to a loopback interface. In IPv6 only deployments, this 32-bit LSR Id must be derived by some other means that guarantees global uniqueness within the MPLS network, similar to that of BGP Identifier [RFC6286].

This document qualifies the first sentence of last paragraph of Section 2.5.2 of [RFC5036] to be per address family and therefore updates that sentence to the following: "For a given address family over which a Hello is sent, and a given label space, an LSR MUST advertise the same transport address." This rightly enables the per-platform label space to be shared between IPv4 and IPv6.

In summary, this document not only allows the usage of a common LDP identifier i.e. same LSR-Id (aka LDP Router-Id), but also the common Label space id for both IPv4 and IPv6 on a dual-stack LSR.

This document reserves 0.0.0.0 as the LSR-Id, and prohibits its usage.

5. Peer Discovery

5.1. Basic Discovery Mechanism

Section 2.4.1 of [RFC5036] defines the Basic Discovery mechanism for directly connected LSRs. Following this mechanism, LSRs periodically sends LDP Link Hellos destined to "all routers on this subnet" group multicast IP address.

Interesting enough, per the IPv6 addressing architecture [RFC4291], IPv6 has three "all routers on this subnet" multicast addresses:

FF01:0:0:0:0:0:0:2 = Interface-local scope

FF02:0:0:0:0:0:0:2 = Link-local scope

FF05:0:0:0:0:0:0:2 = Site-local scope

[RFC5036] does not specify which particular IPv6 'all routers on this subnet' group multicast IP address should be used by LDP Link Hellos.

This document specifies the usage of link-local scope e.g. FF02:0:0:0:0:0:0:2 as the destination multicast IP address in IPv6 LDP Link Hellos. An LDP Hello packet received on any of the other destination addresses must be dropped. Additionally, the link-local IPv6 address MUST be used as the source IP address in IPv6 LDP Link Hellos.

Also, the LDP Link Hello packets must have their IPv6 Hop Limit set to 255, and be checked for the same upon receipt before any further processing, as specified in Generalized TTL Security Mechanism (GTSM)[RFC5082]. The built-in inclusion of GTSM automatically protects IPv6 LDP from off-link attacks.

More importantly, if an interface is a dual-stack LDP interface (e.g. enabled with both IPv4 and IPv6 LDP), then the LSR must periodically send both IPv4 and IPv6 LDP Link Hellos (using the same LDP Identifier per section 4) and must separately maintain the Hello adjacency for IPv4 and IPv6 on that interface.

In summary, the IPv4 and IPv6 LDP Link Hellos must carry the same LDP identifier (assuming per-platform label space usage).

5.2. Extended Discovery Mechanism

Suffice to say, the extended discovery mechanism (defined in section 2.4.2 of [RFC5036]) doesn't require any additional IPv6 specific consideration, since the targeted LDP Hellos are sent to a pre-configured (unicast) destination IPv6 address.

The link-local IP addresses MUST NOT be used as the source or destination IPv6 addresses in extended discovery.

6. LDP Session Establishment and Maintenance

Section 2.5.1 of [RFC5036] defines a two-step process for LDP session establishment, once the peer discovery has completed (LDP Hellos have been exchanged):

1. Transport connection establishment

2. Session initialization

The forthcoming sub-sections discuss the LDP consideration for IPv6 and/or dual-stacking in the context of session establishment and maintenance.

6.1. Transport connection establishment

Section 2.5.2 of [RFC5036] specifies the use of an optional transport address object (TLV) in LDP Link Hello message to convey the transport (IP) address, however, it does not specify the behavior of LDP if both IPv4 and IPv6 transport address objects (TLV) are sent in a Hello message or separate Hello messages. More importantly, it does not specify whether both IPv4 and IPv6 transport connections should be allowed, if there were Hello adjacencies for both IPv4 and IPv6 whether over a single interface or multiple interfaces.

This document specifies that:

1. An LSR MUST NOT send a Hello containing both IPv4 and IPv6 transport address optional objects. In other words, there MUST be at most one optional Transport Address object in a Hello message. An LSR MUST include only the transport address whose address family is the same as that of the IP packet carrying Hello.
2. An LSR SHOULD accept the Hello message that contains both IPv4 and IPv6 transport address optional objects, but MUST use only the transport address whose address family is the same as that of the IP packet carrying Hello.
3. An LSR MUST send separate Hellos (each containing either IPv4 or IPv6 transport address optional object) for each IP address-family, if LDP was enabled for both IP address-families.
4. An LSR MUST use a global unicast IPv6 address in IPv6 transport address optional object of outgoing targeted hellos, and check for the same in incoming targeted hellos (i.e. MUST discard the hello, if it failed the check).
5. An LSR MUST prefer using global unicast IPv6 address for an LDP session with a remote LSR, if it had to choose between global unicast IPv6 address and link-local IPv6 address (pertaining to the same LDP Identifier) for the transport connection.

6. An LSR SHOULD NOT create (or honor the request for creating) a TCP connection for a new LDP session with a remote LSR, if they already have an LDP session (for the same LDP Identifier) established over whatever IP version transport.

This means that only one transport connection is established, even if there are two Hello adjacencies (one for IPv4 and another for IPv6). This is independent of whether the Hello Adjacencies are created over a single interface (scenario 1 in section 1.1) or multiple interfaces (scenario 2 in section 1.1) between two LSRs.

7. An LSR SHOULD prefer the LDP/TCP connection over IPv6 for a new LDP session with a remote LSR, if it has both IPv4 and IPv6 hello adjacencies for the same LDP Identifier (over a dual-stack interface, or two or more single-stack IPv4 and IPv6 interfaces). This applies to the section 2.5.2 of RFC5036.
8. An LSR SHOULD prefer the LDP/TCP connection over IPv6 for a new LDP session with a remote LSR, if they attempted two TCP connections using IPv4 and IPv6 transport addresses simultaneously.

An implementation may provide an option to favor one AFI (IPv4, say) over another AFI (IPv6, say) for the TCP transport connection, so as to use the preferred IP version for the LDP session, and derive deterministic active/passive roles.

6.2. Maintaining Hello Adjacencies

As outlined in section 2.5.5 of RFC5036, this draft describes that if an LSR has a dual-stack interface, which is enabled with both IPv4 and IPv6 LDP, then the LSR must periodically send both IPv4 and IPv6 LDP Link Hellos and must separately maintain the Hello adjacency for IPv4 and IPv6 on that interface.

This ensures successful labeled IPv4 and labeled IPv6 traffic forwarding on a dual-stacked interface, as well as successful LDP peering using the appropriate transport on a multi-access interface (even if there are IPv4-only, IPv6-only and dual-stack LSRs connected to that multi-access interface).

6.3. Maintaining LDP Sessions

Two LSRs maintain a single LDP session between them (i.e. not tear down an existing session), as described in section 6.1, whether

- they are connected via a dual-stack LDP enabled interface or via two single-stack LDP enabled interfaces;
- a single-stack interface is converted to a dual-stack interface (e.g. figure 1) on either LSR;
- an additional single-stack or dual-stack interface is added or removed between two LSRs (e.g. figure 2).

Needless to say that the procedures defined in section 6.1 should result in preferring LDPoIPv6 session only after the loss of an existing LDP session (because of link failure, node failure, reboot etc.).

On the other hand, if a dual-stack interface is converted to a single-stack interface (by disabling IPv4 or IPv6 routing), then the LDP session should be torn down ONLY if the disabled IP version was the same as that of the transport connection. Otherwise, the LDP session should stay intact.

If the LDP session is torn down for whatever reason (LDP disabled for the corresponding transport, hello adjacency expiry etc.), then the LSRs should initiate establishing a new LDP session as per the procedures described in section 6.1 of this document along with RFC5036.

7. Label Distribution

An LSR SHOULD NOT advertise both IPv4 and IPv6 FEC-label bindings (as well as interface addresses via ADDRESS message) from/to the peer over an LDP session (using whatever transport), unless it has valid IPv4 and IPv6 Hello Adjacencies for that peer, as specified in section 6.2.

Another solution for getting the same result as above is by negotiating the IP Capability for a given AFI, as specified in [IPPWCap].

An LSR MUST NOT allocate and advertise FEC-Label bindings for link-local IPv6 address, and ignore such bindings, if ever received. An LSR MUST treat the IPv4-mapped IPv6 address, defined in section

2.5.5.2 of [RFC4291], the same as that of a global IPv6 address and not mix it with the 'corresponding' IPv4 address.

Additionally, to ensure backward compatibility (and interoperability with IPv4-only LDP implementations), this document specifies that -

1. An LSR MUST NOT send a label mapping message with a FEC TLV containing FEC Elements of different address-family. In other words, a FEC TLV in the label mapping message MUST contain the FEC Elements belonging to the same address-family.
2. An LSR MUST NOT send an Address message (or Address Withdraw message) with an Address List TLV containing IP addresses of different address-family. In other words, an Address List TLV in the Address (or Address Withdraw) message MUST contain the addresses belonging to the same address-family.

8. LDP Identifiers and Next Hop Addresses

RFC5036 section 2.7 specifies logic for mapping between a peer LDP Identifier and the peer's addresses to find the correct LIB entry for any prefix by using a database populated by the Address message. However, this logic is insufficient to deal with overlapping IPv6 (link-local) addresses used by two or more peers. One may note that all interior IP routing protocols specify using link-local IPv6 addresses as the next-hops.

This document specifies that the logic is enhanced with the usage of (Hello Adjacency) database populated by the Hello messages. This additional database lookup is useful if/when two or more peers use the same link-local IPv6 address as the IP routing next-hops (causing duplicate next-hop entries).

Specifically, this document specifies that an LSR should (continue to) use the machinery described in RFC5036 section 2.7 to map between a peer LDP Identifier and the peer's addresses (learned via ADDRESS message) for any prefix. However, if this mapping fails (for reasons such as the one described earlier), then an LSR can find the peer LDP Identifier by checking for the particular link-local IPv6 address and interface (corresponding to the next-hop in the unicast routing table) in the hello adjacency database.

If an LSR can't find such a mapping in either database, then LSR should follow procedures specified in RFC5036 (e.g. not resolve the label).

Lastly, for better scale and optimization, an LSR may advertise only the link-local IPv6 addresses in the Address message, assuming that the peer uses only the link-local IPv6 addresses as static and/or dynamic IP routing next-hops.

9. LDP TTL Security

This document also specifies that the LDP/TCP transport connection over IPv6 (i.e. LDPoIPv6) must follow the Generalized TTL Security Mechanism (GTSM) procedures (Section 3 of [RFC5082]) for an LDP session peering established between the adjacent LSRs using Basic Discovery, by default.

In other words, GTSM is enabled by default for an IPv6 LDP peering session using Basic Discovery. This means that the 'IP Hop Limit' in IPv6 packet is set to 255 upon sending, and checked to be 255 upon receipt. The IPv6 packet must be dropped failing such a check upon receipt.

The reason GTSM is enabled for Basic Discovery by default, but not for Extended Discovery is that the usage of Basic Discovery typically results in a single-hop LDP peering session, whereas the usage of Extended Discovery typically results in a multi-hop LDP peering session. While the latter is deemed out of scope (section 1.2), in line with GTSM [RFC5082], it is worth clarifying the following exceptions that may occur with Basic or Extended Discovery usage:

- a) Two adjacent LSRs (i.e. back-to-back PE routers) forming a single-hop LDP peering session after doing an Extended Discovery (for Pseudowire, say)
- b) Two adjacent LSRs forming a multi-hop LDP peering session after doing a Basic Discovery, due to the way IP routing changes between them (temporarily (e.g. session protection) or permanently)
- c) Two adjacent LSRs (i.e. back-to-back PE routers) forming a single-hop LDP peering session after doing both Basic and Extended Discovery

In (a), GTSM is not enabled for the LDP peering session by default, hence, it would not do any harm or good.

In (b), GTSM is enabled by default for the LDP peering session by default and enforced, hence, it would prohibit the LDP peering session from getting established.

In (c), GTSM is enabled by default for Basic Discovery and enforced on the subsequent LDP peering. However, if each LSR uses the same IPv6 transport address object value in both Basic and Extended discoveries, then it would result in a single LDP peering session and that would be enabled with GTSM. Otherwise, GTSM would not be enforced on the 2nd LDP peering session corresponding to the Extended Discovery.

This document allows for the implementation to provide an option to statically (configuration) and/or dynamically override the default behavior (enable/disable GTSM) on a per-peer basis. This would also address the exception (b) above. Suffice to say that such an option could be set on either LSR (since GTSM negotiation would ultimately disable GTSM between LSR and its peer(s)).

The built-in GTSM inclusion is intended to automatically protect IPv6 LDP peering session from off-link attacks.

10. IANA Considerations

None.

11. Security Considerations

The extensions defined in this document only clarify the behavior of LDP, they do not define any new protocol procedures. Hence, this document does not add any new security issues to LDP.

While the security issues relevant for the [RFC5036] are relevant for this document as well, this document reduces the chances of off-link attacks when using IPv6 transport connection by including the use of GTSM procedures [RFC5082].

Moreover, this document allows the use of IPsec [RFC4301] for IPv6 protection, hence, LDP can benefit from the additional security as specified in [RFC4835] as well as [RFC5920].

12. Acknowledgments

We acknowledge the authors of [RFC5036], since the text in this document is borrowed from [RFC5036].

Thanks to Bob Thomas for providing critical feedback to improve this document early on. Thanks to Eric Rosen, Lizhong Jin, Bin Mo, Mach Chen, and Kishore Tiruveedhula for reviewing this document. The authors also acknowledge the help of Manoj Dutta and Vividh Siddha.

Also, thanks to Andre Pelletier (who brought up the issue about active/passive determination, and helped us craft the appropriate solutions.

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

13. Additional Contributors

The following individuals contributed to this document:

Kamran Raza
Cisco Systems, Inc.
2000 Innovation Drive
Kanata, ON K2K-3E8, Canada
Email: skraza@cisco.com

Nagendra Kumar
Cisco Systems, Inc.
SEZ Unit, Cessna Business Park,
Bangalore, KT, India
Email: naikumar@cisco.com

14. References

14.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC4291] Hinden, R. and S. Deering, "Internet Protocol Version 6 (IPv6) Addressing Architecture", RFC 4291, February 2006.
- [RFC5036] Andersson, L., Minei, I., and Thomas, B., "LDP Specification", RFC 5036, October 2007.
- [RFC5082] Pignataro, C., Gill, V., Heasley, J., Meyer, D., and Savola, P., "The Generalized TTL Security Mechanism (GTSM)", RFC 5082, October 2007.

14.2. Informative References

- [RFC4301] Kent, S. and K. Seo, "Security Architecture and Internet Protocol", RFC 4301, December 2005.
- [RFC4835] Manral, V., "Cryptographic Algorithm Implementation Requirements for Encapsulating Security Payload (ESP) and Authentication Header (AH)", RFC 4835, April 2007.
- [RFC5920] Fang, L., "Security Framework for MPLS and GMPLS Networks", RFC 5920, July 2010.
- [RFC4798] De Clercq, et al., "Connecting IPv6 Islands over IPv4 MPLS Using IPv6 Provider Edge Routers (6PE)", RFC 4798, February 2007.
- [IPPWCap] Raza, K., "LDP IP and PW Capability", draft-ietf-mpls-ldp-ip-pw-capability, June 2011.

Author's Addresses

Vishwas Manral
Hewlett-Packard, Inc.
19111 Pruneridge Ave., Cupertino, CA, 95014
Phone: 408-447-1497
Email: vishwas.manral@hp.com

Rajiv Papneja
Huawei Technologies
2330 Central Expressway
Santa Clara, CA 95050
Phone: +1 571 926 8593
EMail: rajiv.papneja@huawei.com

Rajiv Asati
Cisco Systems, Inc.
7025 Kit Creek Road
Research Triangle Park, NC 27709-4987
Email: rajiva@cisco.com

Carlos Pignataro
Cisco Systems, Inc.
7200 Kit Creek Road
Research Triangle Park, NC 27709-4987
Email: cpignata@cisco.com

Internet Engineering Task Force
Internet-Draft
Intended status: Standards Track
Expires: December 16, 2012

Q. Zhao
Huawei Technology
L. Fang
C. Zhou
Cisco Systems
L. Li
China Mobile
N. So
Verizon Business
K. Kamran
Cisco Systems
July 16, 2012

LDP Extensions for Multi Topology Routing
draft-ietf-mpls-ldp-multi-topology-04.txt

Abstract

Multi-Topology (MT) routing is supported in IP networks with the use of MT aware IGP protocols. In order to provide MT routing within Multiprotocol Label Switching (MPLS) Label Distribution Protocol (LDP) networks new extensions are required.

This document describes the LDP protocol extensions required to support MT routing in an MPLS environment.

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 16, 2012.

Copyright Notice

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

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

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

Table of Contents

1. Terminology.....	3
2. Introduction.....	3
3. Signaling Extensions.....	4
3.1. Topology-Scoped FEC.....	4
3.2. New Address Families: MT IP.....	4
3.3. LDP FEC Elements with MT IP AF.....	5
3.4. IGP MT-ID Mapping and Translation.....	6
3.5. LDP MT Capability Advertisement.....	6
3.6. Procedures.....	7
3.7. LDP Sessions.....	8
3.8. Reserved MT ID Values.....	9
4. MT Applicability on FEC-based features.....	9
4.1. Typed Wildcard FEC Element.....	9
4.2. End-of-LIB.....	10
5. Error Handling.....	10
5.1. MT Error Notification "Invalid Topology ID".....	10
6. Backwards Compatibility.....	10
7. MPLS Forwarding in MT.....	10
8. Security Consideration.....	10
9. IANA Considerations.....	11
10. Acknowledgement.....	12
11. Contributors' Addresses.....	12
12. References.....	13

12.1. Normative References.....	13
12.2. Informative References.....	13
Authors' Addresses.....	13
Appendix.....	14
A. Requirements.....	14
B. Application Scenarios.....	15
B.1. Simplified Data-plane.....	15
B.2. Using MT for P2P Protection.....	15
B.3. Using MT for mLDP Protection.....	16
B.4. Service Separation.....	16
B.5. An Alternative inter-AS VPN Solution.....	16

1. Terminology

This document uses MPLS terminology defined in [RFC5036]. Additional terms are defined below:

- o MT-ID: A 16 bit value used to represent the Multi-Topology ID.
- o Default MT Topology: A topology that is built using the MT-ID default value of 0.
- o MT Topology: A topology that is built using the corresponding MT-ID.

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. Introduction

Multi-Topology (MT) routing is supported in IP networks with the use of MT aware IGP protocols. It would be advantageous for communications Service Providers (CSP) to support Multiple Topologies (MT) within MPLS environments (MPLS-MT). Beneficial MPLS-MT deployment applications include:

- o A CSP may want to assign varying QoS profiles to traffic, based on a specific MT.
- o Separate routing and MPLS domains may be used to isolated multicast and IPv6 islands within the backbone network.
- o Specific IP address space could be routed across an MT based on security or operational isolation requirements.
- o Low latency links could be assigned to an MT for delay sensitive traffic.

- o Management traffic could be separated from customer traffic using multiple MTs, where the management traffic MT does not use links that carries customer traffic.

This document describes the LDP procedures and protocol extensions required to support MT routing in an MPLS environment.

3. Signaling Extensions

3.1. Topology-Scoped FEC

LDP assigns and binds a label to a FEC, where a FEC is a list of one or more FEC elements. To setup LSPs for unicast IP routing paths, LDP assigns local labels for IP prefixes, and advertises these labels to its peers so that an LSP is setup along the routing path. To setup MT LSPs for all IP prefixes under a given topology scope, the LDP "prefix-related" FEC element must be extended to include topology info. This infers that MT-ID becomes an attribute of Prefix-related FEC element, and all FEC-Label binding operations are performed under the context of given topology (MT-ID).

The following Subsection (3.2 New Address Families: MT IP) defines the extension required to bind "prefix-related" FEC to a topology.

3.2. New Address Families: MT IP

The LDP base specification [RFC5036] (Section 4.1) defines the "Prefix" FEC Element as follows:

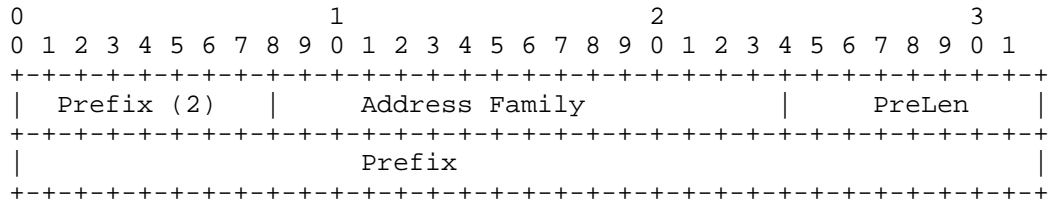


Figure 1: Prefix FEC Element Format [RFC5036]

Where "Prefix" encoding is as defined for given "Address Family", and whose length (in bits) is specified by the "PreLen" field.

To extend IP address families for MT, two new Address Families named "MT IP" and "MT IPv6" are used to specify IPv4 and IPv6 prefixes within a topology scope.

The format of data associated with these new Address Family is:

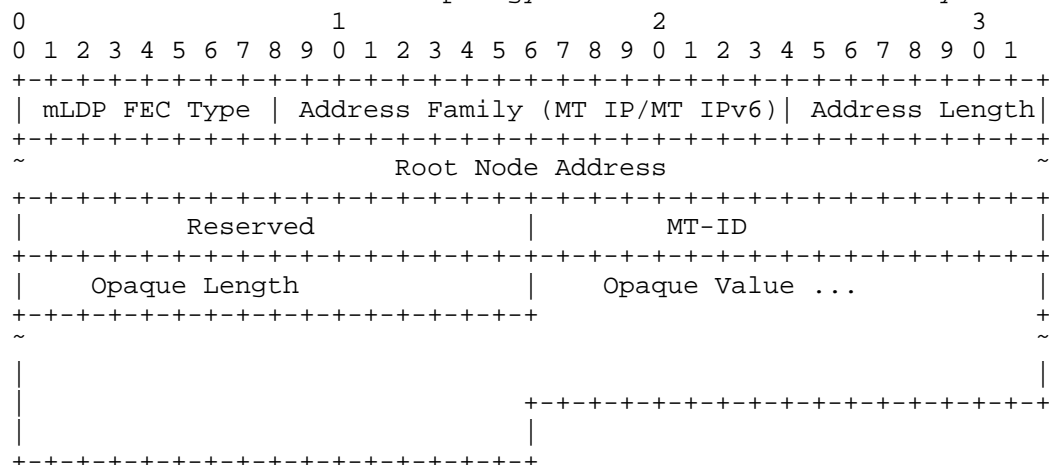


Figure 4: MT mLDP FEC Element Format

The MT Typed Wildcard FEC element encoding is as follows:

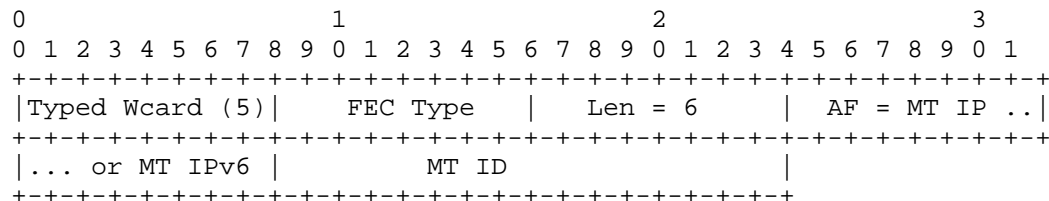


Figure 5: MT Typed Wildcard FEC Element

3.4. IGP MT-ID Mapping and Translation

The non-reserved non-special IGP MT-ID values can be used/carried in LDP as-is and need no translation. However, there is a need for translating reserved/special IGP MT-ID values to corresponding LDP MT-IDs. The corresponding special/reserved LDP MT-ID values are defined in later section 10.

3.5. LDP MT Capability Advertisement

We specify a new LDP capability, named "Multi-Topology (MT)", which is defined in accordance with LDP Capability definition guidelines [RFC5561]. The LDP "MT" capability can be advertised by an LDP speaker to its peers either during the LDP session initialization or after the LDP session is setup to announce LSR capability to support MTR for the given IP address family.

The "MT" capability is specified using "Multi-Topology Capability" TLV. The "Multi-Topology Capability" TLV format is in accordance with LDP capability guidelines as defined in [RFC5561]. To be able to specify IP address family, the capability specific data (i.e. "Capability Data" field of Capability TLV) is populated using "Typed Wildcard FEC Element" as defined in [RFC5918].

The format of "Multi-Topology Capability" TLV is as follows:

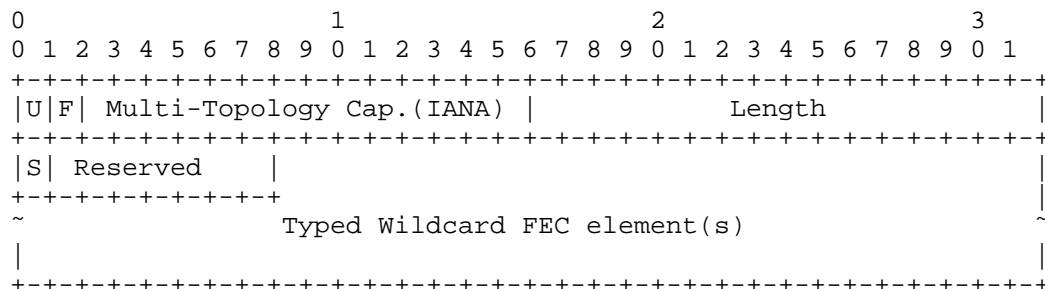


Figure 6: Multi-Topology Capability TLV Format

Where:

- o U- and F-bits: MUST be 1 and 0, respectively, as per Section 3 of LDP Capabilities [RFC5561].
- o Multi-Topology Capability: Capability TLV type (IANA assigned)
- o S-bit: MUST be 1 if used in LDP "Initialization" message. MAY be set to 0 or 1 in dynamic "Capability" message to advertise or withdraw the capability respectively.
- o Typed Wildcard FEC element(s): One or more elements specified as the "Capability data".
- o Length: The length (in octets) of TLV.

The encoding of Typed Wcard FEC element, as defined in [RFC5561], is defined in the Section 3.3 of this document.

3.6. Procedures

To announce its MT capability for an IP address family, LDP FEC type, and Multi Topology, an LDP speaker MAY send "MT Capability" including the exact Typed Wildcard FEC element with corresponding "Address Family" field (i.e. set to "MT IP" for IPv4 and set to "MT IPv6" for IPv6 address family), corresponding "FEC

Type" field (i.e. set to "P2P", "P2MP", "MP2MP"), and corresponding "MT-ID". To announce its MT capability for both IPv4 and IPv6 address family, or for multiple FEC types, or for multiple Multi Topologies, an LDP speaker MAY send "MT Capability" with one or more MT Typed FEC elements in it.

- o The capability for supporting multi-topology in LDP can be advertised during LDP session initialization stage by including the LDP MT capability TLV in LDP Initialization message. After an LDP session is established, the MT capability can also be advertised or withdrawn using Capability message (only if "Dynamic Announcement" capability [RFC5561] has already been successfully negotiated).
- o If an LSR has not advertised MT capability, its peer must not send messages that include MT identifier to this LSR.
- o If an LSR receives a Label Mapping message with MT parameter from downstream LSR-D and its upstream LSR-U has not advertised MT capability, an LSP for the MT will not be established.
- o We propose to add a new notification event to signal the upstream that the downstream is not capable.
- o If an LSR is changed from non-MT capable to MT capable, it sets the S bit in MT capability TLV and advertises via the Capability message. The existing LSP is treated as LSP for default MT (ID 0).
- o If an LSR is changed from LDP-MT capable to non-MT capable, it may initiate withdraw of all label mapping for existing LSPs of all non-default MTs. Then it clears the S bit in MT capability TLV and advertises via the Capability message.
- o If an LSR is changed from IGP-MT capable to non-MT capable, it may wait until the routes update to withdraw FEC and release the label mapping for existing LSPs of specific MT.

3.7. LDP Sessions

Depending on the number of label spaces supported, if a single global label space is supported, there will be one session supported for each pair of peer, even there are multiple topologies supported between these two peers. If there are different label spaces supported for different topologies, which means that label spaces overlap with each other for different MTs, then it is suggested to establish multiple sessions for multiple topologies between these two peers. In this case, multiple LSR-IDs need to be allocated beforehand so that each multiple topology can have its own label space ID.

3.8. Reserved MT ID Values

Certain MT topologies are assigned to serve predetermined purposes:

Default-MT: Default topology. This corresponds to OSPF default IPv4 and IPv6, as well as ISIS default IPv4. A value of 0 is proposed.

ISIS IPv6 MT: ISIS default MT-ID for IPv6.

Wildcard-MT: This corresponds to All-Topologies. A value of 65535 (0xffff) is proposed.

We propose a new IANA registry "LDP Multi-Topology ID Name Space" under IANA "LDP Parameter" namespace to keep LDP MT-ID reserved value.

If an LSR receives a FEC element with an "MT-ID" value that is "Reserved" for future use (and not IANA allocated yet), the LSR must abort the processing of the FEC element, and SHOULD send a notification message with status code "Invalid MT-ID" to the sender.

4. MT Applicability on FEC-based features

4.1. Typed Wildcard FEC Element

[RFC5918] extends base LDP and defines Typed Wildcard FEC Element framework. Typed Wildcard FEC element can be used in any LDP message to specify a wildcard operation/action for given type of FEC.

The MT extensions proposed in document do not require any extension in procedures for Typed Wildcard FEC element, and these procedures apply as-is to MT wildcarding. The MT extensions, though, allow use of "MT IP" or "MT IPv6" in the Address Family field of the Typed Wildcard FEC element in order to use wildcard operations in the context of a given topology. The use of MT-scoped address family also allows us to specify MT-ID in these operations.

The proposed format in section 4.3 allows an LSR to perform wildcard FEC operations under the scope of a topology. If an LSR wishes to perform wildcard operation that applies to all topologies, it can use "Wildcard Topology" MT-ID as defined in section 4.8. For instance, upon local un-configuration of topology "x", an LSR may send wildcard label withdraw with MT-ID "x" to withdraw all its labels from peer that were advertised under the scope of topology "x". On the other hand, upon some global configuration change, an LSR may send wildcard label withdraw with MT-ID set to "Wildcard Topology" to withdraw all its labels under all topologies from the peer.

[RFC5919] specifies extensions and procedures for an LDP speaker to signal its convergence for given FEC type towards a peer. The procedures defined in [RFC5919] apply as-is to MT FEC element. This means that an LDP speaker MAY signal its IP convergence using Typed Wildcard FEC element, and its MT IP convergence per topology using MT Typed Wildcard FEC element (as defined in earlier section).

5. Error Handling

The extensions defined in this document utilise the existing LDP error handling defined in [RFC5036]. If an LSR receives an error notification from a peer for an MPLS-MT session, it terminates the LDP session by closing the TCP transport connection for the session and discarding all MT-ID label mappings learned via the session.

5.1. MT Error Notification "Invalid Topology ID"

If an LSR has advertized an MT Capability TLV using the Initialization message or Capability message, which includes Typed Wildcard FEC Element(s) with specific MT-ID(s), and it receives an MT message with a MT-ID which is not included in the supported list, it should response this "Invalid Topology ID" status code.

6. Backwards Compatibility

The MPLS-MT solution is backwards compatible with existing LDP enhancements defined in [RFC5036], including message authenticity, integrity of message, and topology loop detection.

7. MPLS Forwarding in MT

Although forwarding is out of the scope of this draft, we include some forwarding consideration for informational purpose here.

The specified signaling mechanisms allow all the topologies to share the platform-specific label space; this is the feature that allows the existing data plane techniques to be used; and the specified signaling mechanisms do not provide any way for the data plane to associate a given packet with a context-specific label space.

8. Security Consideration

No specific security issues with the proposed solutions are known. The proposed extension in this document does not introduce any new

9. IANA Considerations

The document introduces following new protocol elements that require IANA consideration and assignments:

- o New LDP Capability TLV: "Multi-Topology Capability" TLV (requested code point: 0x510 from LDP registry "TLV Type Name Space").
- o New Status Code: "Multi-Topology Capability not supported" (requested code point: 0x50 from LDP registry "Status Code Name Space").
- o New Status Code: "Invalid Topology ID" (requested code point: 0x51 from LDP registry "Status Code Name Space").
- o New Status Code: "Unknown Address Family" (requested code point: 0x52 from LDP registry "Status Code Name Space").

Registry:		
Range/Value	E	Description
-----	---	-----
0x00000051	1	Invalid Topology ID

Figure 7: New Status Codes for LDP Multi Topology Extensions

- o New address families under IANA registry "Address Family Numbers":
 - MT IP: Multi-Topology IP version 4 (requested codepoint: 26)
 - MT IPv6: Multi-Topology IP version 6 (requested codepoint: 27)

Figure 8: Address Family Numbers

- o New registry "LDP Multi-Topology (MT) ID Name Space" under "LDP Parameter" namespace. The registry is defined as:

Range/Value	Name
-----	-----
0	Default Topology (ISIS and OSPF)
1-4095	Unassigned
4096	ISIS IPv6 routing topology (i.e. ISIS MT ID #2)
4097-65534	Reserved (for future allocation)
65535	Wildcard Topology (ISIS or OSPF)

Figure 9: LDP Multi-Topology (MT) ID Name Space

10. Acknowledgement

The authors would like to thank Dan Tappan, Nabil Bitar, Huang Xin, Eric Rosen, IJsbrand Wijnands, Dimitri Papadimitriou, Yiqun Chai for their valuable comments on this draft.

11. Contributors' Addresses

Raveendra Torvi
Juniper Networks
10, Technoogy Park Drive
Westford, MA 01886-3140
US

Email: rtorvi@juniper.net

Huaimo Chen
Huawei Technology
125 Nagog Technology Park
Acton, MA 01719
US

Email: huaimochen@huawei.com

Emily Chen
Huawei Technology
2330 Central Expressway
Santa Clara, CA 95050
US

Email: emily.chenying@huawei.com

Chen Li
China Mobile
53A, Xibianmennei Ave.
Xunwu District, Beijing 01719
China

Email: lichenyj@chinamobile.com

Lu Huang
China Mobile
53A, Xibianmennei Ave.
Xunwu District, Beijing 01719
China

Email: huanglu@chinamobile.com

Email: E-mail: daniel@olddog.co.uk

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.
- [RFC5036] Andersson, L., Minei, I., and B. Thomas, "LDP Specification", RFC 5036, October 2007.
- [RFC5561] Thomas, B., Raza, K., Aggarwal, S., Aggarwal, R., and JL. Le Roux, "LDP Capabilities", RFC 5561, July 2009.
- [RFC5918] Asati, R., Minei, I., and B. Thomas, "Label Distribution Protocol (LDP) 'Typed Wildcard' Forward Equivalence Class (FEC)", RFC 5918, August 2010.
- [RFC5919] Asati, R., Mohapatra, P., Chen, E., and B. Thomas, "Signaling LDP Label Advertisement Completion", RFC 5919, August 2010.

12.2. Informative References

- [RFC5920] Fang, L., "Security Framework for MPLS and GMPLS Networks", RFC 5920, July 2010.

Authors' Addresses

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

Email: quintin.zhao@huawei.com

Luyuan Fang
Cisco Systems
300 Beaver Brook Road
Boxborough, MA 01719
US

Email: lufang@cisco.com

Chao Zhou
Cisco Systems
300 Beaver Brook Road
Boxborough, MA 01719
US

Email: czhou@cisco.com

Lianyuan Li
China Mobile
53A, Xibianmennei Ave.
Xunwu District, Beijing 01719
China

Email: lilianyuan@chinamobile.com

Ning So
Verizon Business
2400 North Glenville Drive
Richardson, TX 75082
USA

Email: Ning.So@verizonbusiness.com

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

Email: E-mail: skraza@cisco.com

Appendix A. Requirements

The following specific requirements and objectives have been defined in order to provide the functionality described in Section 2 (Introduction), and facilitate CSP configuration and operation:

- o Minimise configuration and operation complexity of MPLS-MT across the network.
- o The MPLS-MT solution SHOULD NOT require data-plane modification.

- o The MPLS-MT solution MUST support multiple topologies. Allowing an MPLS LSP to be established across a specific, or set of, multiple topologies.
- o Control and filtering of LSPs using explicitly including or excluding multiple topologies MUST be supported.
- o The MPLS-MT solution MUST be capable of supporting QoS mechanisms.
- o The MPLS-MT solution MUST be backwards compatibility with existing LDP message authenticity and integrity techniques, and loop detection.
- o Deployment of MPLS-MT within existing MPLS networks should be possible, with nodes not capable of MPLS-MT being unaffected.

Appendix B. Application Scenarios

B.1. Simplified Data-plane

IGP-MT requires additional data-plane resources maintain multiple forwarding for each configured MT. On the other hand, MPLS-MT does not change the data-plane system architecture, if an IGP-MT is mapped to an MPLS-MT. In case MPLS-MT, incoming label value itself can determine an MT, and hence it requires a single NHLFE space. MPLS-MT requires only MT-RIBs in the control-plane, no need to have MT-FIBs. Forwarding IP packets over a particular MT requires either configuration or some external means at every node, to map an attribute of incoming IP packet header to IGP-MT, which is additional overhead for network management. Whereas, MPLS-MT mapping is required only at the ingress-PE of an MPLS-MT LSP, because of each node identifies MPLS-MT LSP switching based on incoming label, hence no additional configuration is required at every node.

B.2. Using MT for P2P Protection

Mechanisms exist that can configure alternate path via backup-mt, such that if primary link fails, then backup-MT can be used for forwarding. However, such techniques require special marking of IP packets that needs to be forwarded using backup-MT. MPLS-LDP-MT procedures simplify the forwarding of the MPLS packets over backup-MT, as MPLS-LDP-MT procedure distribute separate labels for each MT. How backup paths are computed depends on the implementation, and the algorithm. The MPLS-LDP-MT in conjunction with IGP-MT could be used to separate the primary traffic and backup traffic. For example, service providers can create a backup MT that consists of links that are meant only for backup traffic. Service providers can then establish bypass LSPs, standby LSPs, using backup MT, thus keeping undeterministic backup traffic away from the primary traffic.

B.3. Using MT for mLDP Protection

For the P2MP or MP2MP LSPs setup by using mLDP protocol, there is a need to setup a backup LSP to have an end to end protection for the primary LSP in the applications such as IPTV, where the end to end protection is a must. Since the mLDP LSP is setup following the IGP routes, the second LSP setup by following the IGP routes can not be guaranteed to have the link and node diversity from the primary LSP. By using MPLS-LDP-MT, two topology can be configured with complete link and node diversity, where the primary and secondary LSP can be set up independently within each topology. The two LSPs setup by this mechanism can protect each other end-to-end.

B.4. Service Separation

MPLS-MT procedures allow establishing two distinct LSPs for the same FEC, by advertising separate label mapping for each configured topology. Service providers can implement QoS using MPLS-MT procedures without requiring to create separate FEC address for each class. MPLS-MT can also be used separate multicast and unicast traffic.

B.5. An Alternative inter-AS VPN Solution

When the LSP is crossing multiple domains for the inter-as VPN scenarios, the LSP setup process can be done by configuring a set of routers which are in different domains into a new single domain with a new topology ID using the LDP multiple topology. All the routers belong this new topology will be used to carry the traffic across multiple domains and since they are in a single domain with the new topology ID, so the LDP LSP set up can be done without propagating VPN routes across AS boundaries.

Network Working Group
Internet-Draft
Intended status: Informational
Expires: October 31, 2012

Y. Weingarten

S. Bryant
Cisco
N. Sprecher
Nokia Siemens Networks
D. Ceccarelli
D. Caviglia
F. Fondelli
Ericsson
M. Corsi
Altran
B. Wu
X. Dai
ZTE Corporation
April 29, 2012

Applicability of MPLS-TP Linear Protection for Ring Topologies
draft-ietf-mpls-tp-ring-protection-02.txt

Abstract

This document presents an applicability statement to address the requirements for protection of ring topologies for Multi-Protocol Label Switching Transport Profile (MPLS-TP) Label Switched Paths (LSP) on multiple layers. The MPLS-TP Requirements document specifies specific criteria for justification of dedicated protection mechanism for particular topologies, including optimizing the number of OAM entities needed, minimizing the number of labels for protection paths, minimizing the number of recovery elements in the network, and minimizing the number of control and management transactions necessary. The document proposes a methodology for ring protection based on existing MPLS-TP survivability mechanisms, specifically those defined in MPLS-TP Linear Protection, without the need for specification of new constructs or protocols.

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

Status of this Memo

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

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

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

This Internet-Draft will expire on October 31, 2012.

Copyright Notice

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

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

Table of Contents

1. Introduction	4
1.1. Problem statement	4
1.2. Terminology and Notation	5
1.3. Contributing Authors	7
2. P2P Ring Protection	7
2.1. Wrapping	7
2.2. Steering	9
2.3. P2P ring protection using SPME	10
2.3.1. Path SPME for Steering	11
2.3.2. Wrapping with segment based SPME	12
2.3.3. Wrapping node protection	13
2.3.4. Wrapping for link and node protection	14
2.4. Analysis of p2p protection	14
3. P2MP protection	15
3.1. Wrapping for p2mp LSP	15
3.1.1. Comparison of Wrapping and ROM-Wrapping	17
3.1.2. Multiple Failures Comparison	19
3.2. Steering for p2mp paths	19
3.2.1. Context labels	20
3.2.2. Walkthrough using context labels	22
4. Coordination protocol	23
5. Conclusions and Recommendations	24
6. IANA Considerations	25
7. Security Considerations	25
8. Acknowledgements	25
9. Informative References	25
Authors' Addresses	26

1. Introduction

Multi-Protocol Label Switching Transport Profile (MPLS-TP) is being standardized as part of a joint effort between the Internet Engineering Task Force (IETF) and the International Telecommunication Union Standardization (ITU-T). These specifications are based on the requirements that were generated from this joint effort.

The requirements for MPLS-TP [TPReqs] indicates a requirement to support a network that may include sub-networks that constitute a MPLS-TP ring as defined in the requirements. The requirements document does not identify any protection requirements specific to a ring topology. However, the requirements state that specific protection mechanisms aimed at ring topologies may be developed if these allow the network to optimize:

- o Number of OAM entities needed to trigger the protection
- o Number of elements of recovery needed
- o Number of labels required
- o Number of control and management plane transactions during a recovery operation
- o Impact of signaling and routing information exchanged, in presence of control plane

This document will propose a set of basic mechanisms that could be used for the protection of the data flows that traverse a MPLS-TP ring. The mechanism is based on existing MPLS and MPLS-TP protection mechanisms. We show that this mechanism provides the ability to protect all of the basic conditions within a reasonable time frame and does optimize the criteria set out in [TPReqs] as summarized above.

A related topic in [TPReqs] addresses the required support for interconnected rings. This topic involves various scenarios that require further study and will be addressed in a separate document, based on the principles outlined in this document.

1.1. Problem statement

Ring topologies, as defined in [TPReqs], are used in transport networks due to their ability to easily support both p2p and p2mp transport paths. When designing a protection mechanism for a ring topology, there is a need to address both -

1. A point-to-point transport path that enters a MPLS-TP capable ring at one node, the ingress node, and exits the ring at a single egress node possibly continuing beyond the ring.
2. Where the ring is being used as a branching point for a point-to-multipoint transport path, i.e. the transport path enters the MPLS-TP capable ring at the ingress node and exits through a number of egress nodes, possibly continuing beyond the ring.

In either of these two situations, there is a need to address the following different cases -

1. One of the ring links causes a fault condition. This could be either a unidirectional or bidirectional fault, and should be detected by the neighboring nodes.
2. One of the ring nodes causes a fault condition. This condition is invariably a bidirectional fault (although in rare cases of misconfiguration this could be detected as a unidirectional fault) and should be detected by the two neighboring ring nodes.
3. An operator command is issued to a specific ring node. A description of the different operator commands is found in Section 4.12 of [RFC4427]. Examples of these commands include Manual Switch, Forced Switch, or Clear operations.

The protection domain addressed in this document is limited to the traffic that is traversing the ring. Traffic on the transport path prior to the ring ingress node or beyond the egress nodes may be protected by some other mechanism.

1.2. Terminology and Notation

The terminology used in this document is based on the terminology defined in the MPLS-TP framework documents:

- o MPLS-TP Framework[TPFwk]
- o MPLS-TP OAM Framework[OAMFwk]
- o MPLS-TP Survivability Framework[SurvivFwk]

The MPLS-TP Framework document [TPFwk] defines a Sub-Path Maintenance Entity (SPME) construct that can be defined between any two LSRs of a MPLS-TP LSP. This SPME may be configured as a co-routed bidirectional path. The SPME is defined to allow management and monitoring of any segment of a transport path. This concept will be used extensively throughout the document to support protection of the

traffic that traverses a MPLS-TP ring.

In addition, we describe the use of the label stack in connection with the redirecting of data packets by the protection mechanism. 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 levels however, we only present the levels that are germane to the protection mechanism.
3. When applicable, the S-bit (signifying that a given label is the bottom of the label stack) will be denoted by the string '+S' within the label. If a label is not shown with '+S' that label may or may not be the bottom label in the stack. '+S' is only shown when it is important to illustrate that a given label is definitely the last one in the label stack.
4. The label of the LSP at the ingress point to the ring will be denoted by the string "LI" and the label of the LSP that is expected at the egress point from the ring will be denoted by the string "LE", and "LSE" will denote the label expected at the exit LSR of a SPME (if it is different from the egress point from the ring).
5. The label for a SPME will be denoted by Pxi(y) where x and y are LSR identifiers and the intention is to the label for LSR-x to transmit to LSR-y over the SPME whose index is i.

For example -

- o the label stack [LI] denotes the label stack received at the ingress node of the ring. This may have additional labels after LI, e.g. a PW label however, this is irrelevant to the discussion of the protection scenario.
- o [PBl(G)|LE] denotes a stack whose top-label is the SPME-1 label for LSR-B to transmit the data packet to LSR-G, the second label is the label that would be used by the egress LSR to continue the packet on the original LSP.
- o If "LE" were the bottom label in the stack, then the label stack would be shown as [PBl(G)|LE+S].

1.3. Contributing Authors

Akira Sakurai (NEC), Rolf Winter (NEC)

2. P2P Ring Protection

Classically there are two protection architecture mechanisms for ring topologies, based on SDH specifications [G.841], that have been proposed in various forums to perform recovery of a topological ring network - "wrapping" and "steering". The following sub-sections will examine these two mechanisms.

2.1. Wrapping

Wrapping is defined as a local protection architecture. This mechanism is local to the LSRs that are neighbors to the detected fault. When a fault is detected (either a link or node failure), the neighboring LSR can identify that the fault would prevent forwarding of the data along the data path. Therefore, in order to continue the data along the path, there is need to "wrap" all data traffic around the ring, on an alternate data path, until arriving at the LSR that is on the opposite side of the fault. When this LSR also detects that there is a fault condition on the LSP, it can identify that the data traffic that is arriving on the alternate (protecting) data path is intended for the "broken" LSP. Therefore, again taking a local decision, can wrap the data back onto the normal working path until the egress from the ring segment. Wrapping behavior is similar to MPLS-TE FRR as defined in [RFC4090] using either bypass or detour tunnels. It would be possible to wrap each LSP around the failed links via a detour tunnel using a different label for each LSP or to wrap all the LSPs using a bypass tunnel and a single label.

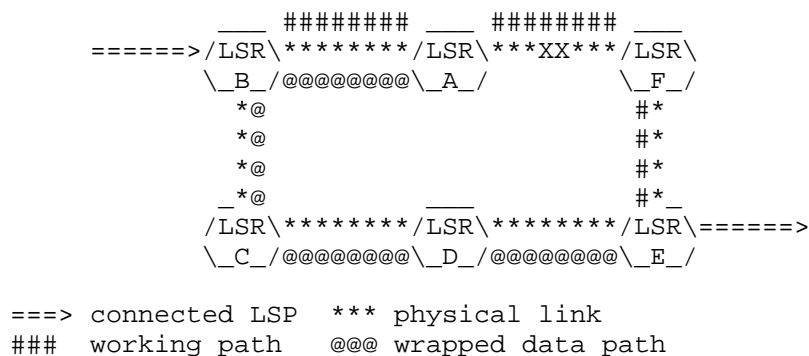


Figure 1: Wrapping protection for p2p path

In this figure we have a ring with a LSP that enters the ring at LSR-B and exits at LSR-E. The normal working path follows through B-A-F-E. If a fault is detected on the link A<-->F, then the wrapping mechanism decides that LSR-A would wrap the traffic around the ring, on a wrapped data path A-B-C-D-E-F, to arrive at LSR-F (on the far side of the failed link). LSR-F would then wrap the data packets back onto the working path F-->E to the egress node. In this protection scheme, the traffic will follow the path - B-A-B-C-D-E-F-E.

This protection scheme is simple in the sense that there is no need for coordination between the different LSR in the ring - only the LSRs that detect the fault must wrap the traffic, either onto the wrapped data path (at the near-end) or back to the working path (at the far-end). Coordination would only be needed to maintain co-routed bidirectional traffic even in cases of a unidirectional fault condition.

The following considerations should be taken into account when considering use of wrapping protection:

- o Detection of loss-of-continuity or mis-connectivity, should be performed at the link level and/or per LSR when using node-level protection. Configuration of the protection being performed (i.e. link protection or node protection) needs to be performed a-priori, since the configuration of the proper protection path is dependent upon this decision.
- o There is a need to define a data-path that traverses the alternate path around the ring to connect between the two neighbors of the detected fault. If protecting both the links and the nodes of a LSP, then, for a ring with N nodes, there is a need for $O(2N)$ alternate paths.
- o When wrapping, the data is transmitted over some of the links twice, once in each direction. For example, in the figure above the traffic is transmitted both B-->A and then A-->B, later it is transmitted E-->F and F-->E. This means that there is additional bandwidth needed for this protection.
- o If a double-fault situation occurs in the ring, then wrapping will not be able to deliver any packets except between the ingress and the first fault location. This is based on the need for wrapping to connect between the neighbors of the fault location, and this is not possible in the segmented ring.
- o The resource allocation for the alternate-paths could be problematic, since most of these alternate paths will not be used

simultaneously. One possibility could be to allocate '0' resources and depend on the NMS to allocate the proper resources around the ring.

- o Wrapping also involves greater latency in delivering the packets, as a result of traversing the entire ring. This could be very restrictive for large rings.

2.2. Steering

The second common scheme for ring protection, steering, takes advantage of the ring topology by defining two paths from the ingress point (to the ring) to the egress point going in opposite directions around the ring. This is illustrated in Figure 2, where if we assume that the traffic needs to enter the ring from node B and exit through node F, we could define a primary path through nodes B-A-F, and an alternate path through the nodes B-C-D-E-F. In steering the switching is always performed by the ingress node (node B in Figure 2). If a fault condition is detected anywhere on the working path (B-A-F), then the traffic would be redirected by B to the alternate path (i.e. B-C-D-E-F).

This mechanism bears similarities to linear 1:1 protection [SurvivFwk]. The two paths around the ring act as the working and protection paths. There is need to communicate to the ingress node the need to switch over to the protection path and there is a need to coordinate the switchover between the two end-points of the protected domain.

The following considerations must be taken into account regarding the steering architecture:

- o Steering relies on a failure detection method that is able to notify the ingress node of the fault condition. This may involve different OAM functionality described in [OAMFwk], e.g. Remote Defect Indication, Alarm reporting.
- o The process of notifying the ingress node adds to the latency of the protection switching process, after the detection of the fault condition.
- o While there is no need for double bandwidth for the data path, there is the necessity for the ring to maintain enough capacity for all of the data in both directions around the ring.

2.3. P2P ring protection using SPME

The SPME concept was introduced by [TPFwk] to support management and monitoring an arbitrary segment of a transport. However, an SPME is essentially a valid LSP that may be used to aggregate all LSP traffic that traverses the sub-path delineated by the SPME. An SPME may be monitored using the OAM mechanisms as described in the MPLS-TP OAM Framework document [OAMFwk].

When defining a MPLS-TP ring as a protection domain, there is a need to design a protection mechanism that protects all the LSPs that cross the MPLS-TP ring. For this purpose, we associate a (working) SPME with the segment of the transport path that traverses the ring. In addition, we configure an alternate (protecting) SPME that traverses the ring in the opposite direction around the ring. The exact selection of the SPMEs is dependent on the type of transport path and protection that is being implemented and will be detailed in the following sub-sections.

Based on this architectural configuration for ring protection, it is possible to limit the number of alternate paths needed to protect the data traversing the ring. In addition, since we will perform all of the OAM functionality on the SPME configured for the traffic, we can minimize the number of OAM sessions needed to monitor the data traffic of the ring - rather than monitoring each individual LSP.

The following figure shows a MPLS-TP ring that is part of a larger MPLS-TP network. The ring could be used as a network segment that may be traversed by numerous LSPs. In particular, the figure shows that for all LSPs that connect to the ring at LSR-B and exit the ring from LSR-F, we configure two SPME through the ring (the first SPME traverses along B-A-F, and the second SPME traverses B-C-D-E-F).

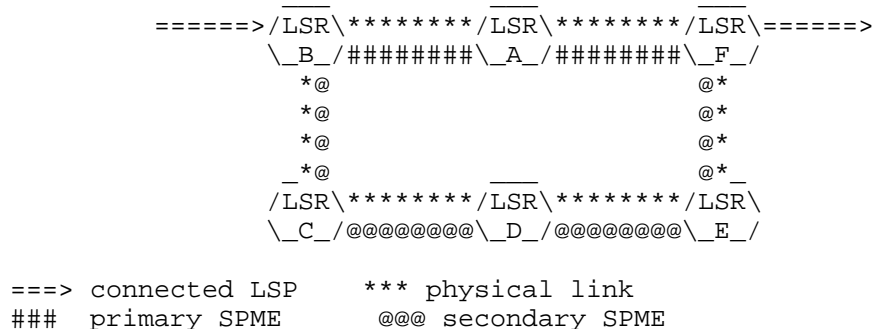


Figure 2: A MPLS-TP ring

In all of the following subsections, we use 1:1 linear protection [SurvivFwk] [LinProtect] to perform protection switching and coordination when a signal fault is detected. The actual configuration of the SPMEs used may change dependent upon the choice of methodology and this will be detailed in the following sections. However, in all of these configurations the mechanism will be to transmit the data traffic on the primary SPME, while applying OAM functionality over both the primary and the secondary SPME to detect signal fault conditions on either path. If a signal fault is detected on the primary SPME, then the mechanism described in [LinProtect] shall be used to coordinate a switch-over of data traffic to the secondary SPME.

Assuming that the SPME is implemented as an hierarchical LSP, packets that arrive at LSR-B with a label stack [LI] will have the SPME label pushed at LSR-B and the LSP label will be swapped for the label that is expected by the egress LSR (i.e. the packet will arrive at LSR-A with a label stack of [PA1(B)|LE], arrive at LSR-F with [PE1(F)|LE]). The SPME label will be popped by LSR-F and the LSP label will be treated appropriately at LSR-F and forwarded along the LSP, outside the ring. This scenario is true for all LSP that are aggregated by this primary SPME.

2.3.1. Path SPME for Steering

A p2p SPME that traverses part of a ring has two Maintenance Entity Group End Points (MEPs), each one acts as the ingress and egress in one direction of the bidirectional SPME. Since the SPME is traversing a ring we can take advantage of another characteristic of a ring - there is always an alternative path between the two MEPs, i.e. traversing the ring in the opposite direction. This alternative SPME can be defined as the protection path for the working path that is configured as part of the LSP and defined as a SPME.

For each pair of SPMEs that are defined in this way, it is possible to verify the connectivity and continuity by applying the MPLS-TP OAM functionality to both the working and protection SPME. If a discontinuity or mis-connectivity is detected then the MEPs will become aware of this condition, and could perform a protection switch of all LSPs to the alternate, protection SPME.

This protection mechanism is identical to application of 1:1 linear protection [SurvivFwk] [LinProtect] to the pair of SPMEs. Under normal conditions, all LSP data traffic will be transmitted on the working SPME. If the linear protection is triggered, by either the OAM indication, an other fault indication trigger, or an operator command, then the MEPs will select the protection SPME to transmit all LSP data packets.

The protection SPME will continue to transmit the data packets until the stable recovery of the fault condition. Upon recovery, the ingress LSR could switch traffic back to the working SPME, if the protection domain is configured for revertive behavior.

The control of the protection switching, especially for cases of operator commands, would be covered by the protocol defined in [LinProtect].

2.3.2. Wrapping with segment based SPME

It is possible to use the SPME mechanism to perform segment-based protection. For each link in the ring, we define two SPME - the first is a SPME between the two LSRs that are connected by the link, and the second SPME between these same two LSRs but traversing the entire ring (except the link that connects the LSRs). In Figure 3 we show the primary SPME that connects LSR-A & LSR-F over a segment connection, and the secondary SPME that connects these same LSRs by traversing the ring in the opposite direction.

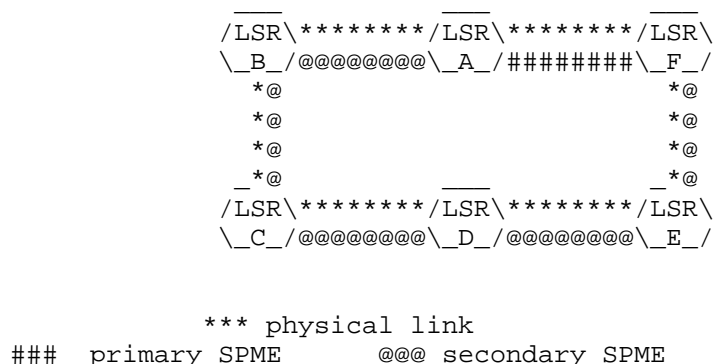


Figure 3: Segment SPMEs

By applying OAM monitoring of these two SPME (at each LSR), it is possible to affect a wrapping protection mechanism for the LSP traffic that traverses the ring. The LSR on either side of the segment would identify that there is a fault condition on the link and redirect all LSP traffic to the secondary SPME. The traffic would traverse the ring until arriving at the neighboring (relative to the segment) LSR. At this point, the LSP traffic would be redirected onto the original LSP, quite likely over the neighboring SPME.

Following the progression of the label stack through this switching operation (for a LSP that enters the ring at LSR B and exits the ring

at LSR E):

1. The data packet arrives at LSR-A with label stack [L1+S] (i.e. top label from the LSP and bottom-of-stack indicator)
2. In the normal case (no switching), LSR-A forwards the packet with label stack [PA1(F)|LSE+S] (i.e. swap the label for the LSP, to be acceptable to the SPME egress, and push the label for the primary SPME from LSR-A to LSR-F).
3. When switching is in-effect, LSR-A forwards the packet with label stack [PA2(B)|LSE+S] (i.e. LSR-A pushed the label for the secondary SPME from LSR-A to LSR-F, after swapping the label of the lower level LSP). This will be transmitted along the secondary SPME until LSR-E forwards it to LSR-F with label stack [PE2(F)|LSE+S].
4. When the packet arrives at LSR-F, it will pop the SPME label, process the LSP label, and forward the packet to the next point, possibly pushing a SPME label if the next segment is likewise protected.

2.3.3. Wrapping node protection

Implementation of protection at the node level would be similar to the mechanism described in the previous sub-section. The difference would be in the SPMEs that are used. For node protection, the primary SPME would be configured between the two LSR that are connected to the node that is being protected (see SPME between LSR-A and LSR-E through LSR-F in Figure 4), and the secondary SPME would be configured between these same nodes, going around the ring (see secondary SPME in Figure 4).

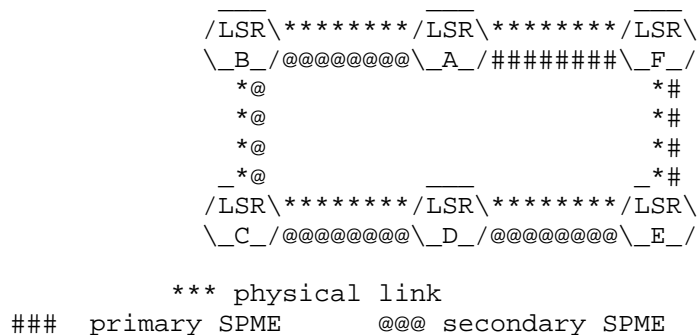


Figure 4: Node-protection SPMEs

The protection mechanism would work similarly - based on 1:1 linear protection [SurvivFwk], triggered by OAM functions on both SPMEs, and wrapping the data packets onto the secondary SPME at the ingress MEP (e.g. LSR-A in the figure) of the SPME and back onto the continuation of the LSP at the egress MEP (e.g. LSR-E in the figure) of the SPME.

2.3.4. Wrapping for link and node protection

In the different types of wrapping presented in Section 2.3.2 and Section 2.3.3, there is a limitation that the protection mechanism must a priori decide whether it is protecting for link or node failure. In addition, the neighboring LSR, that detects the fault, cannot readily differentiate between a link failure or a node failure.

It would be possible to configure extra SPME to protect both for link and node failures, arriving at a configuration of the ring that is shown in Figure 5. Choosing the SPME to use for the wrapping would, however, then involve considerable effort and could result in the protected traffic not sharing the same protection path in both directions.

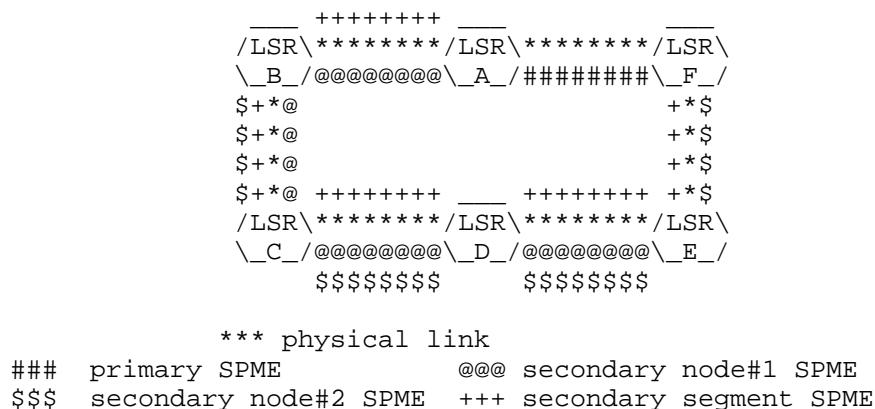


Figure 5: Segment & Node protection SPMEs

2.4. Analysis of p2p protection

Analyzing the mechanisms described in the above subsections we can point to the following observations (based on a ring with N nodes, assumed to be not more than 16):

- o Number of SPME that need to be configured - for steering SPME protection (Section 2.3.1) = $O(2N^2)$ [two SPME from each ingress LSR to each other node in the ring], for wrapping based on SPME either as described in Section 2.3.2 and Section 2.3.3 = $O(2N)$ [however, the operator must decide a priori on whether to protect for link failures or node failures at each point]
- o Number of OAM sessions at each node - for steering = $O(2N)$, for SPME wrapping = 3
- o Bandwidth requirements - for SPME-based steering: single bandwidth at each link, for wrapping: double bandwidth at links that are between ingress and wrapping node and between second wrapping node and egress.
- o Special considerations - for SPME based steering: latency of OAM detection of fault condition by ingress MEP [using Alarm-reporting could optimize over using CC-V only], for SPME wrapping: at each node must decide a priori whether protecting for link or node failures. To protect for both node and link failures would increase the complexity of deciding which protection path to use, as well as, violating the co-routedness of the protected traffic.

Based on this analysis, using steering as described in Section 2.3.1 would be the recommended protection mechanism due to its simplicity, even though it may involve the use of additional resources (i.e. SPME) for monitoring the traffic. It should be pointed out that the number of SPME involved in this protection could be reduced by eliminating SPME between pairs of LSR that are not used as an ingress and egress pair.

3. P2MP protection

[TPReqs] requires that ring protection must provide protection for unidirectional point-to-multipoint paths through the ring. Ring topologies provide a ready platform for supporting such data paths. A p2mp LSP in an MPLS-TP ring would be characterized by a single ingress LSR and multiple egress LSRs. The following sub-sections will present methods to address the protection of the ring-based sections of these LSP.

3.1. Wrapping for p2mp LSP

When protecting a p2mp ring data path using the wrapping architecture, the basic operation is similar to the description given, as the traffic has been wrapped back onto the normal working path on the far-side of the detected fault and will continue to be

transported to all of the egress points.

It is possible to optimize the performance of the wrapping mechanism when applied to p2mp LSPs by exploiting the topology of ring networks.

This improved mechanism, which we call Ring Optimized Multicast Wrapping (ROM-Wrapping), behaves much the same as classical wrapping. There is one difference - rather than configuring the protection LSP between the end nodes of a failed link (link protection) or between the upstream and downstream node of a failed node (node protection), the improved mechanism configures a protection p2mp LSP from the upstream (with respect to the failure) node and all egress nodes (for the particular LSP) downstream from the failure.

Referring to Figure 6, it is possible to identify the protected (working) LSP (A-B-[C]-[D]-E-[F]) and one possible backup (protection) LSP. This protection LSP will be used to wrap the data back around the ring to protect against a failure on link B-C. This protection LSP is also a p2mp LSP that is configured with egress points (at nodes F, D, & C) complimentary to the broken working data path.

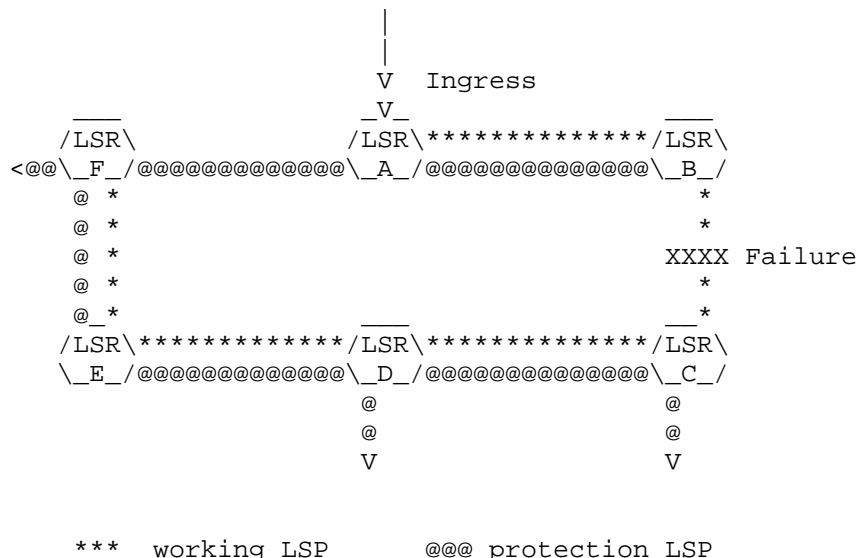


Figure 6: P2MP ROM Wrapping

Using this mechanism, there is a need to configure a particular protection LSP for each node on the working LSP. In the table below, "X's Backup" is the backup path activated by node X as a consequence

of a failure affecting node Y (downstream node with respect to X) or link X-Y, and square brackets, in the path, indicate egress nodes.

Protected LSP: A->B->[C]->[D]->E->[F]

----- LINK/NODE PROTECTION-----

A's Backup:	A->[F]->E->[D]->[C]
B's Backup:	B->A->[F]->E->[D]->[C]
C's Backup:	C->B->A->[F]->E->[D]
D's Backup:	D->C->B->A->[F]
E's Backup:	E->D->C->B->A->[F]

It should be noted that ROM-Wrapping is an LSP based protection mechanism, as opposed to the SPME based protection mechanisms that are presented in other sections of this draft. While this may seem to be limited in scope, the mechanism may be very efficient for many applications that are based on p2mp distribution schemes. While ROM-Wrapping can be applied to any network topology, it is particularly efficient for interconnected ring topologies.

3.1.1. Comparison of Wrapping and ROM-Wrapping

It is possible to compare the Wrapping and the ROM-Wrapping mechanisms in different aspects, and show some improvements offered by ROM-Wrapping.

When configuring the protection LSP for Wrapping it is necessary to configure for a specific failure: link protection or node protection. If the protection method is configured to protect node failures but the actual failure affects a link, this could result in failing to deliver traffic to the node, when it should be possible to.

ROM-Wrapping however does not have this limitation, because there is no distinction between node and link protection. Whether link B-C or node C fails, in either case the rerouting will attempt to reach C. If the failure is on the link, the traffic will be delivered to C, while if the failure is at node C, the traffic will be rerouted correctly until node D, and will be blocked at this point. However, all egress nodes up-to the failure will be able to deliver the traffic properly.

A second aspect is the number of hops needed to properly deliver the traffic. Referring to the example shown in Figure 6, where a failure is detected on link B-C, the following table lists the set of nodes traversed by the data in the protection:

Basic Wrapping:

A-B	B-A-F-E-D-C	[C]-[D]-E-[F]
"Upstream" segment	backup path	"Downstream" segment
with respect to the		with respect to the
failure		failure

ROM Wrapping:

A-B	B-A-[F]-E-[D]-[C]	..
"Upstream" segment	backup path	
with respect to the		
failure		

Comparing the two lists of nodes, it is possible to see that in this particular case the number of hops crossed using the simple Wrapping is significantly higher than the number of hops crossed by the traffic when ROM-Wrapping is used. Generally, the number of hops for basic Wrapping is always higher or at least equal compared to ROM-Wrapping. This implies a certain waste of bandwidth on all links that are crossed in both directions.

Considering the ring network previously seen, it is possible to do some bandwidth utilization considerations. The protected LSP is set up from A to F clockwise and an M Mbps bandwidth is reserved along the path. All the protection LSPs are pre-provisioned counterclockwise, each of them may also have reserved bandwidth M. These LSPs share the same bandwidth in a SE (Shared Explicit) [RSVP] style.

The bandwidth reserved counterclockwise is not used when the protected LSP is properly working and could, in theory, be used for extra traffic [RFC4427]. However, it should be noted that [TPReqs] does not require support of such extra traffic.

The two recovery mechanism require different protection bandwidths. In the case of Wrapping, the bandwidth used is M in both directions of many of the links. While in case of ROM-Wrapping, only the links from the ingress node to the node performing the actual wrapping utilize M bandwidth in both directions, while all other links utilize M bandwidth only in the counterclockwise direction.

Consider the case of a failure detected on link B-C as shown in Figure 6. The following table lists the bandwidth utilization on each link (in units equal to M), for each recovery mechanism and for each direction (CW=clockwise, CCW=counterclockwise).

	Wrapping	ROM-Wrapping
Link A-B	CW+CCW	CW+CCW
Link A-F	CCW	CCW
Link F-E	CW+CCW	CCW
Link E-D	CW+CCW	CCW
Link D-C	CW+CCW	CCW

3.1.2. Multiple Failures Comparison

A further comparison between Wrapping and ROM-Wrapping can be done with respect to their ability to react to multiple failures. The wrapping recovery mechanism does not have the ability to recover from multiple failures on a ring network, while ROM-Wrapping is able to recover, from some multiple failures.

Consider, for example, a double link failure affecting links B-C and C-D shown in Figure 6. The Wrapping mechanism is not able to recover from the failure because B, upon detecting the failure, has no alternative paths to reach C. The whole P2MP traffic is lost. The ROM-Wrapping mechanism is able to partially recover from the failure, because the backup P2MP LSP to node F and node D is correctly set up and continues delivering traffic.

3.2. Steering for p2mp paths

When protecting p2mp traffic that uses a MPLS-TP ring as its branching point, i.e. it enters the ring at a head-end node and exits the ring at multiple nodes, we can employ a steering mechanism based on 1+1 linear protection [SurvivFwk]. We can configure two p2mp unidirectional SPME from each node on the ring that traverse the ring in both directions. These SPME will be configured with an egress at each ring node. In order to be able to properly direct the LSP traffic to the proper egress point for that particular LSP, we need to employ context labeling as defined in [RFC5331]. The method for using these labels is expanded in section 3.2.1.

For every LSP that enters the ring at a given node the traffic will be sent through both of these SPME, each with its own context label and the context-specific label for the particular LSP. The egress nodes should select the traffic that is arriving on the working SPME. In case there is a failure condition, the egress nodes should select the traffic from whichever of the SPME that is arrives at that node, i.e. since one of the two (presumably the working SPME) will be blocked by the failure. In this way, all egress nodes are able to receive the data traffic. While each node detects that there is

connectivity from the ingress point, it continues to select the data that is coming from the working SPME. If a particular node stops receiving the connectivity messages from the working SPME, it identifies that it must switch its selector to read the data packets from the protection SPME.

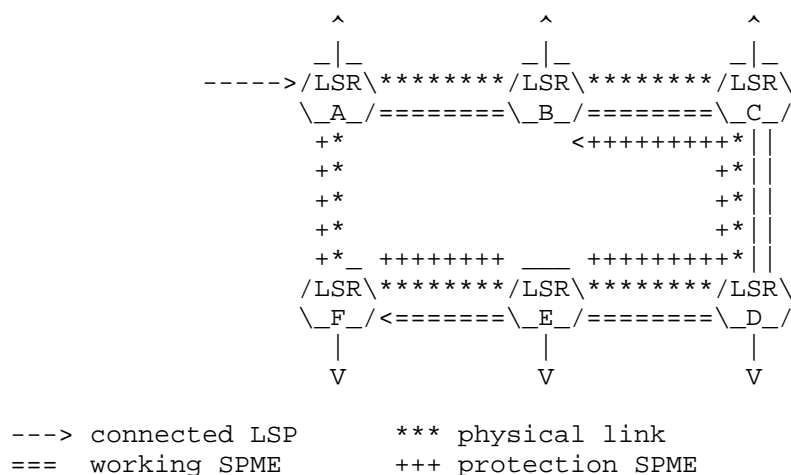


Figure 7: P2MP SPMEs

3.2.1. Context labels

Figure 7 shows the two unidirectional p2mp SPME that are configured from LSR-A with egress points at all of the nodes on the ring. The clockwise SPME (i.e. A-B-C-D-E-F) is configured as the working SPME, that will aggregate all traffic for p2mp LSPs that enter the ring at LSR-A and must be sent out of the ring at any subset of the ring nodes. The counter-clockwise SPME (i.e. A-F-E-D-C-B) is configured as the protection SPME.

[RFC5331] defines the concept of context labels. A context-identifying label defines a context label space that is used to interpret the context-specific labels (found directly below the context- identifying label) for a specific tunnel. The SPME label is a context- identifying label. This means that at each hop the node that receives the SPME label uses it to point not directly to a forwarding table, but to a LIB. As a node receives an SPME label it examines it, discovers that it is a context label, pops off the SPME label, and looks up the next label down in the stack in the LIB indicated by the context label.

The label below this context-identifying label should be used by the forwarding function of the node to decide the actions taken for this

packet. In MPLS-TP ring protection there are two context LIBs. One is the context LIB for the working SPME and the other is the context LIB for the P-SPME. All context LIBs have a behavior defined for the e2e LSP label but the behavior at each node may be different in the context of each SPME.

For example, using the ring that is shown in Figure 7, if the working SPME is configured to have a context-identifying label of CW at each node on the ring and the protection SPME is configured to have a context-identifying label of CP at each node. For the specific LSP we will designate the context-specific label used on the working SPME as WL(x-y) to be the label used as node-x to forward the packet to node-y. Similarly, for the context-specific labels on the protection SPME would be designated PL(x-y). An explicit example of label values appears in the next sub-section.

If we apply the 1+1 linear protection scheme outlined above for an p2mp LSP that enters the ring at LSR-A and has egress points from the ring at LSR-C and LSR-E using the two SPME shown in Figure 7 then a packet that arrives at LSR-A with a label stack [LI+S] will be forwarded on the working SPME with a label stack [CW | WL(A-B)]. The packet should then be forwarded to LSR-C arriving with a label [CW | WL(B-C)], where WL(B-C) should instruct the forwarding function to egress the packet with [LE(C)] and forward a copy to LSR-D with label stack [CW | WL(C-D)].

If a fault condition is detected, then some of the nodes will cease to receive the packets from the clockwise (working) SPME. These LSR should then begin to switch their selector bridge to accept the data packets from the protection SPME. At the ingress point the packet will be transmitted on both the working SPME and the protection SPME. Continuing the example, if there is a failure on the link between LSR-C and LSR-D then LSR-A will transmit one copy of the data to LSR-B with stack [CW | WL(A-B)] and one copy to LSR-F with stack [CP | PL(A-F)]. The packet will arrive at LSR-C from the working SPME and egress from the ring. LSR-E will receive the packet from the protection SPME with stack [CP | PL(F-E)] and the context-sensitive label PL(F-E) will instruct the forwarding function to send a copy out of the ring with label LE(E) and a second copy to LSR-D with stack [CP | PL(E-D)]. In this way each of the egress points receive the packet from the SPME that is available at that point.

This architecture has the added advantages that there is no need for the ingress node to identify the existence of the mis-connectivity, and there is no need for a return path from the egress points to the ingress.

3.2.2. Walkthrough using context labels

In order to better demonstrate the use of the context labels we present a walkthrough of an example application of the p2mp protection presented in this section. Referring to Figure 8, there is a p2mp LSP that traverses the ring, entering the ring at LSR-B and branching off at LSR-D, LSR-E, and LSR-H and does not continue beyond LSR-H. For purposes of protection two p2mp unidirectional SPME are configured on the ring starting from LSR-B. One of the SPME, the working SPME, is configured with egress points at each of the LSR - C, D, E, F, G, H, J, K, A. The second SPME, the protection SPME, is configured with egress points at each of the LSR - A, K, J, H, G, F, E, D, C.

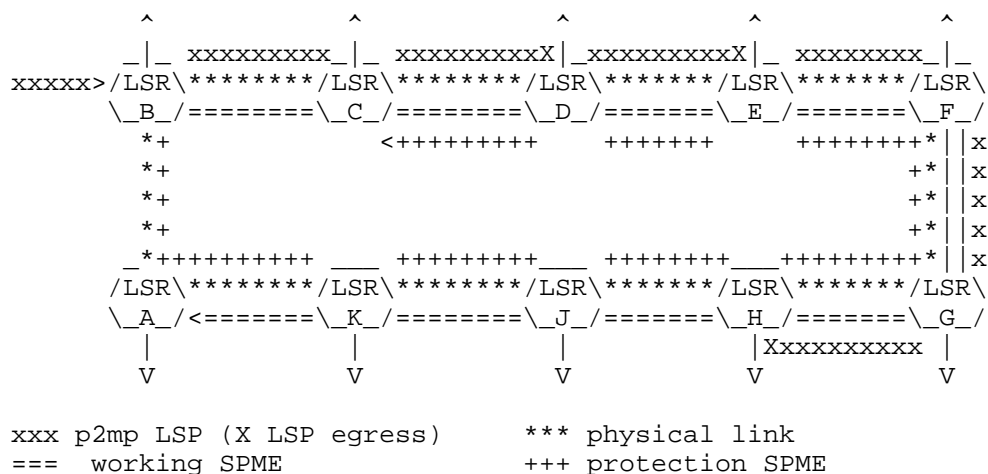


Figure 8: P2MP SPMEs

For this example we suppose that the LSP traffic enters the ring at LSR-B with the label stack [99], leaves the ring at LSR-D with stack [199], at LSR-E with stack [299], and LSR-H with stack [399].

While it is possible for the context-identifying label for the SPME be configured as a different value at each LSR, for the sake of this example we will suppose a configuration of 200 as the context-identifying label for the working SPME at each of the LSR in the ring, and 400 as the context-identifying label for the protection SPME at each LSR.

For the specific connected LSP we configure the following context-specific labels for each context:

node	W-context(200)	P-context(400)
A	65 {drop packet}	165 {fwrdd w/[400 190]}
C	90 {fwrdd w/[200 80]}	190 {drop packet}
D	80 {fwrdd w/[200 75] + egress w/[199]}	180 {egress w/[199]}
E	75 {fwrdd w/[200 65] + egress w/[299]}	175 {fwrdd w/[400 180] + egress w/[299]}
F	65 {fwrdd w/[200 55]}	165 {fwrdd w/[400 175]}
G	55 {fwrdd w/[200 45]}	155 {fwrdd w/[400 165]}
H	45 {egress w/[399]}	145 {fwrdd w/[400 155] + egress w/[399]}
J	65 {drop packet}	165 {fwrdd w/[400 145]}
K	65 {drop packet}	190 {fwrdd w/[400 165]}

When a packet arrives on the LSP to LSR-B with stack [99], the forwarding function determines that it is necessary to forward the packet to both the working SPME with stack [200|90] and the protection SPME with stack [400|165]. Each LSR on the SPME will identify the top label, i.e. 200 or 400, to be the context-identifying label and use the next label in the stack to select the forwarding action from the specific context table.

Therefore, at LSR-C the packet on the working SPME will arrive with stack [200|90] and the 200 will point to the table in the middle column above. After popping the 200 the next label, i.e. 90, will select the forwarding action "fwrdd w/[200|80]" and the packet will be forwarded to LSR-D with stack [200|80]. In this manner, the packet will be forwarded along both SPME according to the configured behavior in the context tables. However, the egress points at LSR D, E, & H, will all be configured with a selector bridge to only use the input from the working SPME. If any of these egress points identify that there is a connection fault on the working SPME, then the selector bridge will cause the LSR to read the input from the protection SPME.

4. Coordination protocol

The Survivability Framework [SurvivFwk] indicates that there is a need to coordinate protection switching between the end-points of a protected bidirectional domain. The coordination is necessary for particular cases, in order to maintain the co-routed nature of the

bidirectional transport path. The particular cases where this becomes necessary include cases of unidirectional fault detection and use of operator commands.

By using the same mechanisms defined in [LinProtect], for linear protection, to apply for ring protection we are able to gain a consistent solution for this coordination between the end-points of the protection domain. The Protection State Coordination Protocol that is specified in [LinProtect] provides coverage for all the coordination cases, including support for operator commands, e.g. Forced-Switch.

5. Conclusions and Recommendations

Ring topologies are prevalent in traditional transport networks and will continue to be used for various reasons. Protection for transport paths that traverse a ring within a MPLS network can be provided by applying an appropriate instance of linear protection, as defined in [SurvivFwk]. This document has shown that for each of the traditional ring protection architectures there is an application of linear protection that provides efficient coverage, based on the use of the Sub-Path Maintenance Entity (SPME), defined in [TPFwk] and [OAMFwk]. For example,

- o p2p Steering - Configuration of two SPME, from ring ingress to ring egress, and 1:1 linear protection
- o p2p Wrapping for link protection - Configuration of two SPME, one for the protected link and the second using the long route between the two neighboring nodes, and 1:1 linear protection.
- o p2p Wrapping for node protection - Configuration of two SPME, one between the two neighbors of the protected node and the second between these two nodes on the long route, and 1:1 linear protection.
- o p2mp Wrapping - it is possible to optimize the performance of the wrapping by configuring the proper protection path to egress the data at the proper branching nodes.
- o p2mp Steering - by combining 1+1 linear protection and configuration of the SPME based on context-sensitive labeling of the protection path.

It has been shown that this set of protection architecture and mechanisms are optimized based on the criteria defined in [TPReqs] for justification of designing a specific protection mechanism for a

ring topology. This thereby alleviates the necessity to create a new mechanism or protocol to support the protection of ring topologies.

By basing the simple p2p ring protection on basic 1:1 linear protection there is a very efficient way of implementing Steering protection for the sections of a transport path that traverses the ring. Steering should be the preferred mechanism for ring protection since it reduces the extra bandwidth required when traffic doubles through wrapped protection, and the ability to protect both against link and node failures without complicating the fault detection or the need to configure multiple protection paths. While this is true, the possibility remains to support either mechanism while depending upon the OAM functionality [outlined in [OAMFwk] and specified in various documents] and the coordination protocol specified for linear protection in [LinProtect].

6. IANA Considerations

This document makes no request of IANA.

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

7. Security Considerations

To be added in future version.

8. Acknowledgements

The authors would like to thank all members of the teams (the Joint Working Team, the MPLS Interoperability Design Team in IETF and the T-MPLS Ad Hoc Group in ITU-T) involved in the definition and specification of MPLS Transport Profile.

9. Informative References

- [RFC4090] Pan, P., Swallow, G., and A. Atlas, "Fast Reroute Extensions to RSVP-TE for LSP Tunnels", RFC 4090, May 2005.
- [RFC5331] Aggarwal, R., Rekhter, Y., and E. Rosen, "MPLS Upstream Label Assignment and Context-Specific Label Space", RFC 5331, Aug 2008.

- [TPReqs] Niven-Jenkins, B., Nadeau, T., and C. Pignataro, "Requirements for the Transport Profile of MPLS", RFC 5654, April 2009.
- [TPFwk] Bocci, M., Bryant, S., Frost, D., and L. Levrau, "MPLS-TP Framework", RFC 5921, May 2010.
- [OAMFwk] Niven-Jenkins, B., Allan, D., and I. Busi, "MPLS-TP OAM Framework", RFC 6371, May 2010.
- [SurvivFwk] Sprecher, N. and A. Farrel, "MPLS-TP Survivability Framework", RFC 6372, June 2010.
- [LinProtect] Sprecher, N., Bryant, S., van Helvoort, H., Fulignoli, A., and Y. Weingarten, "MPLS-TP Linear Protection", RFC 6378, October 2009.
- [RSVP] Braden, R., Zhang, L., Berson, S., Herzog, S., and S. Jamin, "Resource ReSerVation Protocol (RSVP) - Functional Specifications", RFC 2205, September 1997.
- [RFC4427] Mannie, E. and D. Papadimitriou, "Recovery (Protection and Restoration) Terminology for GMPLS", RFC 4427, March 2006.
- [G.841] ITU, "Types and characteristics of SDH network protection architectures", ITU-T G.841, October 1998.

Authors' Addresses

Yaacov Weingarten
34 Hagefen St.
Karnei Shomron, 4485500
Israel

Phone:
Email: wyaacov@gmail.com

Stewart Bryant
Cisco
United Kingdom

Email: stbryant@cisco.com

Nurit Sprecher
Nokia Siemens Networks
3 Hanagar St. Neve Ne'eman B
Hod Hasharon, 45241
Israel

Email: nurit.sprecher@nsn.com

Danielle Ceccarelli
Ericsson
Via A. Negrone 1/A
Genova, Sestri Ponente
Italy

Email: danielle.ceccarelli@ericsson.com

Diego Caviglia
Ericsson
Via A. Negrone 1/A
Genova, Sestri Ponente
Italy

Email: diego.caviglia@ericsson.com

Francesco Fondelli
Ericsson
Via A. Negrone 1/A
Genova, Sestri Ponente
Italy

Email: francesco.fondelli@ericsson.com

Marco Corsi
Altran
Via A. Negrone 1/A
Genova, Sestri Ponente
Italy

Email: corsi.marco@gmail.com

Bo Wu
ZTE Corporation
4F,RD Building 2,Zijinghua Road
Nanjing, Yuhuatai District
P.R.China

Email: wu.bo@zte.com.cn

Xuehui Dai
ZTE Corporation
4F,RD Building 2,Zijinghua Road
Nanjing, Yuhuatai District
P.R.China

Email: dai.xuehui@zte.com.cn

MPLS Working Group
Internet-Draft
Intended status: Standards Track
Expires: December 29, 2012

Chen. Li
Lianyuan. Li
Lu. Huang
China Mobile
Emily. Chen
Quintin. Zhao
Huawei Technologies
June 27, 2012

Management Information Base for MPLS LDP Multi Topology
draft-li-mpls-ldp-mt-mib-03

Abstract

This memo defines an portion of the Management Information Base (MIB) for use with network management protocols in the Internet community. In particular, it describes a MIB module for Multi-Topology Networks over Multi-protocol Label Switching(MPLS) Label Switching Routers(LSRs).

Status of this Memo

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

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

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

This Internet-Draft will expire on December 29, 2012.

Copyright Notice

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

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

to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

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

Table of Contents

1. Introduction	4
2. The Internet-Standard Management Framework	4
3. Overview of MPLS-LDP-MT-STD-MIB objects	4
3.1. MPLS LDP MT Entity Table	4
3.2. MPLS LDP MT Entity Statistics Table	5
3.3. MPLS LDP MT Session Table	5
3.4. MPLS LDP MT In-segment Tables	5
3.5. MPLS LDP MT Out-segment Tables	5
3.6. MPLS LDP MT LSP Table	5
3.7. MPLS LDP MT Notifications	5
4. MPLS-LDP-MT-STD-MIB Module Definitions	6
5. Security Considerations	28
6. IANA Considerations	28
7. Normative References	28
Authors' Addresses	29

1. Introduction

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

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

2. The Internet-Standard Management Framework

For a detailed overview of the documents that describe the current Internet-Standard Management Framework, please refer to section 7 of RFC 3410[RFC3410]. Managed objects are accessed via a virtual information store, termed the Management Information Base or MIB. MIB objects are generally accessed through the Simple Network Management Protocol (SNMP). Objects in the MIB are defined using the mechanisms defined in the Structure of Management Information (SMI). This memo specifies a MIB module that is compliant to the SMIv2, which is described in STD 58, RFC 2578[RFC2578], STD 58, RFC 2579[RFC2579] and STD 58, RFC 2580[RFC2580].

3. Overview of MPLS-LDP-MT-STD-MIB objects

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

3.1. MPLS LDP MT Entity Table

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

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

3.2. MPLS LDP MT Entity Statistics Table

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

3.3. MPLS LDP MT Session Table

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

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

3.4. MPLS LDP MT In-segment Tables

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

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

3.5. MPLS LDP MT Out-segment Tables

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

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

3.6. MPLS LDP MT LSP Table

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

3.7. MPLS LDP MT Notifications

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

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

```

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

    IMPORTS
        IndexIntegerNextFree, IndexInteger
            FROM DIFFSERV-MIB
        InetAddress, InetAddressPrefixLength
            FROM INET-ADDRESS-MIB
        MplsIndexType
            FROM MPLS-LSR-STD-MIB
        MplsLdpLabelType, MplsLspType, MplsLdpIdentifier
            FROM MPLS-TC-STD-MIB
        OBJECT-GROUP, MODULE-COMPLIANCE, NOTIFICATION-GROUP
            FROM SNMPv2-CONF
        transmission, TimeTicks, Integer32, Unsigned32, Counter32
            FROM SNMPv2-SMI
        Counter64, OBJECT-TYPE, MODULE-IDENTITY, NOTIFICATION-TYPE
            FROM SNMPv2-TC;

    mplsLdpMtStdMIB MODULE-IDENTITY
        LAST-UPDATED "201206131436Z"
        ORGANIZATION
            "Multiprotocol Label Switching (mpls) Working Group"
        CONTACT-INFO
            "Chen Li (lichenyj@chinamobile.com)
            Lianyuan Li (lilianyuan@chinamobile.com)
            Lu Huang (huanglu@chinamobile.com)
            China Mobile

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

--
-- Node definitions
--

```



```
-- 1.3.6.1.2.1.10.1.1
```

```
mplsStdMIB OBJECT IDENTIFIER ::= { transmission 166 }
```

```
mplsLdpMtNotifications OBJECT IDENTIFIER ::= { mplsLdpMtStdMIB 0
```

```
}
```

```
mplsLdpMtLspUp NOTIFICATION-TYPE
```

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

```
  STATUS current
```

```
  DESCRIPTION
```

"This notification is generated when the mplsLdpMtLspOperStatus object for one or more contiguous entries in mplsLdpMtLspTable are about to enter the up(1) state from some other state. The included values of mplsLdpMtLspOperSta

```
tus
```

MUST both be set equal to this new state (i.e: up

```
(1)).
```

The two instances of mplsLdpMtLspOperStatus in this notification indicate the range of indexes that are affected. Note that all the indexes of the two ends of the range can be derived from the instance identifiers of these two objects. For cases where a contiguous range of cross-connects have transitioned into the up(1) state at roughly the same time, the device SHOULD issue a single notification for each range of contiguous indexes in an effort to minimize the emission of a large number of notifications. If a notification has to be issued for just a single cross-connect entr

```
y,
```

then the instance identifier (and values) of the two mplsLdpMtLspOperStatus objects MUST be the identical."

```
 ::= { mplsLdpMtNotifications 1 }
```

```
mplsLdpMtLspDown NOTIFICATION-TYPE
```

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

```
  STATUS current
```

```
  DESCRIPTION
```

"This notification is generated when the mplsLdpMtLspOperStatus object for one or more contiguous entries in mplsLdpMtLspTable are about to enter the down(2) state from some other state. The included values of mplsLdpMtLspOperStatus

MUST both be set equal to this down(2) state.
 The two instances of mplsLdpMtLspOperStatus in the
 notification indicate the range of indexes that
 are affected. Note that all the indexes of the two
 ends of the range can be derived from the instance
 identifiers of these two objects. For cases where
 a contiguous range of cross-connects have transitioned
 into the down(2) state at roughly the same time,
 the device SHOULD issue a single notification for each
 range of contiguous indexes in an effort to minimize
 the emission of a large number of notifications.

If a notification has to be issued for just a single
 cross-connect entry, then the instance identifier
 (and values) of the two mplsLdpMtLspOperStatus objects

MUST be identical."
 ::= { mplsLdpMtNotifications 2 }

mplsLdpMtObjects OBJECT IDENTIFIER ::= { mplsLdpMtStdMIB 1 }

mplsLdpMtEntityObjects OBJECT IDENTIFIER ::= { mplsLdpMtObjects 1 }

mplsLdpMtEntityLastChange OBJECT-TYPE

SYNTAX TimeStamp

MAX-ACCESS read-only

STATUS current

DESCRIPTION

"The value of sysUpTime at the time of the most
 recent addition or deletion of an entry
 to/from the mplsLdpMtEntityTable, or
 the most recent change in value of any objects in
 the mplsLdpMtEntityTable.

If no such changes have occurred since the last
 re-initialization of the local management subsystem,

then this object contains a zero value."
 ::= { mplsLdpMtEntityObjects 1 }

mplsLdpMtEntityIndexNext OBJECT-TYPE

SYNTAX IndexIntegerNextFree

MAX-ACCESS read-only

STATUS current

DESCRIPTION

"This object contains an appropriate value to
be used for mplsLdpEntityIndex when creating

Li, et al.

Expires December 29, 2012

[Page 8]

```

        entries in the mplsLdpEntityTable. The value
        0 indicates that no unassigned entries are
        available."
    ::= { mplsLdpMtEntityObjects 2 }

-- mplsLdpMtEntityTable
    mplsLdpMtEntityTable OBJECT-TYPE
        SYNTAX SEQUENCE OF MplsLdpMtEntityEntry
        MAX-ACCESS not-accessible
        STATUS current
        DESCRIPTION
            "This table contains information about the
            MPLS Label Distribution Protocol Multi Topology
            Entities which exist on this Label Switching
            Router (LSR) or Label Edge Router (LER)."
    ::= { mplsLdpMtEntityObjects 3 }

    mplsLdpMtEntityEntry OBJECT-TYPE
        SYNTAX MplsLdpMtEntityEntry
        MAX-ACCESS not-accessible
        STATUS current
        DESCRIPTION
            "An entry in this table represents an LDP MT
            entity. An entry can be created by a network
            administrator or by an SNMP agent as instructed
            by LDP."
        INDEX { mplsLdpMtEntityLdpId, mplsLdpMtEntityMtId, mplsLd
pMtEntityIndex }
    ::= { mplsLdpMtEntityTable 1 }

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

```

```

mplsLdpMtEntityLdpId OBJECT-TYPE
    SYNTAX MplsLdpIdentifier
    MAX-ACCESS not-accessible
    STATUS current
    DESCRIPTION
        "The LDP identifier."
    REFERENCE
        "RFC 5036, LDP Specification, Section on LDP Identifiers."
    ::= { mplsLdpMtEntityEntry 1 }

mplsLdpMtEntityMtId OBJECT-TYPE
    SYNTAX Unsigned32 (0..65535)
    MAX-ACCESS read-create
    STATUS current
    DESCRIPTION
        "The Multi Topology identifier of this LDP MT Entity."
    REFERENCE
        "draft-ietf-mpls-ldp-multi-topology, LDP Extensions for Multi Topology Routing, Section on Multi-Topology ID."
    ::= { mplsLdpMtEntityEntry 2 }

mplsLdpMtEntityIndex OBJECT-TYPE
    SYNTAX IndexInteger
    MAX-ACCESS not-accessible
    STATUS current
    DESCRIPTION
        "This index is used as a secondary index to uniquely identify this row. Before creating a row in this table, the 'mplsLdpMtEntityIndexNext' object should be retrieved. That value should be used for the value of this index when creating a row in this table. NOTE: if a value of zero (0) is retrieved, that indicates that no rows can be created in this table at this time."
    ::= { mplsLdpMtEntityEntry 3 }

mplsLdpMtEntityAdminStatus OBJECT-TYPE
    SYNTAX INTEGER
        {
            enable(1),
            disable(2)
        }
    MAX-ACCESS read-create
    STATUS current
    DESCRIPTION
        "The administrative status of this LDP MT Entity.
    If

```


and
 ntact
 Session
 needs
 he network
 entry
 e
 ed to
 o 'enable',
 ew MT Session."

this object is changed from 'enable' to 'disable'
 this entity has already attempted to establish co
 with a MT Session, then all contact with that MT
 is lost and all information from that MT Session
 to be removed from the MIB. (This implies that t
 management subsystem should clean up any related
 in the mplsLdpMtSessionTable.). At this point th
 operator is able to change values which are relat
 this entity. When the admin status is set back t
 then this MT Entity will attempt to establish a n

```
DEFVAL { enable }
::= { mplsLdpMtEntityEntry 4 }
```

mplsLdpMtEntityStorageType OBJECT-TYPE
 SYNTAX StorageType
 MAX-ACCESS read-create
 STATUS current
 DESCRIPTION
 "The storage type for this conceptual row. Conce
 ptual rows having
 ess to any columnar
 the value 'permanent(4)' need not allow write-acc
 objects in the row."
 ::= { mplsLdpMtEntityEntry 5 }

mplsLdpMtEntityRowStatus OBJECT-TYPE
 SYNTAX RowStatus
 MAX-ACCESS read-create
 STATUS current
 DESCRIPTION
 "The status of this conceptual row. All writable
 objects in this row
 d in detail in the
 Establishment', and
 as been initiated
 wreak havoc with the
 the recommended
 s to down, thereby
 may be modified at any time, however, as describe
 section entitled, 'Changing Values After Session
 again described in the DESCRIPTION clause of the
 mplsLdpMtEntityAdminStatus object, if a session h
 with a Peer, changing objects in this table will
 session and interrupt traffic. To repeat again:
 procedure is to set the mplsLdpMtEntityAdminStatu
 explicitly causing a session to be torn down. Th

en, change objects in this entry, then set the mplsLdpMtEntityAdminS
tatus to enable, which enables a new session to be initiated."
 ::= { mplsLdpMtEntityEntry 6 }

```
-- mplsLdpMtEntityStatsTable
    mplsLdpMtEntityStatsTable OBJECT-TYPE
        SYNTAX SEQUENCE OF MplsLdpMtEntityStatsEntry
        MAX-ACCESS not-accessible
```

```

STATUS current
DESCRIPTION
    "This table contains statistical information for
    LDP MT entities to an LSR."
::= { mplsLdpMtEntityObjects 4 }

mplsLdpMtEntityStatsEntry OBJECT-TYPE
SYNTAX MplsLdpMtEntityStatsEntry
MAX-ACCESS not-accessible
STATUS current
DESCRIPTION
    "An entry in this table is created by the LSR for
    every interface capable of supporting MPLS LDP Mu
lti
ityEntry
tries
ifEntry's
    Topology. It is an extension to the mplsLdpMtEnt
table. Note that the discontinuity behavior of en
tries
in this table MUST be based on the corresponding
    ifDiscontinuityTime."
AUGMENTS { mplsLdpMtEntityEntry }
::= { mplsLdpMtEntityStatsTable 1 }

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

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

```



```

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

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

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

mplsLdpMtEntityStatsHCOctets OBJECT-TYPE
    SYNTAX Counter64
    MAX-ACCESS read-only
    STATUS current
    DESCRIPTION
        "The total number of octets received. This is the
        64 bit version
        of mplsLdpMtEntityStatsOctets, if mplsLdpMtEntity
StatsHCOctets
        is supported according to the rules spelled out i
n RFC2863."
    ::= { mplsLdpMtEntityStatsEntry 5 }

mplsLdpMtEntityStatsDiscontinuityTime OBJECT-TYPE
    SYNTAX TimeTicks
    MAX-ACCESS read-only
    STATUS current
    DESCRIPTION
        "The value of sysUpTime on the most recent occasi
on at which

```



```

any one or more of this MT interface's Counter32
or Counter64
suffered a discontinuity. If no such discontinuit
ies have occurred
since the last re-initialization of the local man
agement subsystem,
then this object contains a zero value."
::= { mplsLdpMtEntityStatsEntry 6 }

mplsLdpMtSessionObjects OBJECT IDENTIFIER ::= { mplsLdpMtObjects
2 }

mplsLdpMtSessionLastChange OBJECT-TYPE
SYNTAX TimeStamp
MAX-ACCESS read-only
STATUS current
DESCRIPTION
    "The value of sysUpTime at the time of the most
    recent addition or deletion to/from the
    mplsLdpMtSessionTable."
::= { mplsLdpMtSessionObjects 1 }

-- mplsLdpMtSessionTable
mplsLdpMtSessionTable OBJECT-TYPE
SYNTAX SEQUENCE OF MplsLdpMtSessionEntry
MAX-ACCESS not-accessible
STATUS current
DESCRIPTION
    "A table of MT Sessions between the LDP MT Entiti
es. Each row in
    this table represents a single MT session."
::= { mplsLdpMtSessionObjects 2 }

mplsLdpMtSessionEntry OBJECT-TYPE
SYNTAX MplsLdpMtSessionEntry
MAX-ACCESS not-accessible
STATUS current
DESCRIPTION
    "An entry in this table represents information on
a single MT
session. The information contained in a row is r
ead-only."
INDEX { mplsLdpMtEntityLdpId, mplsLdpMtEntityMtId, mplsLd
pMtEntityIndex,
        mplsLdpMtSessionPeerId }
::= { mplsLdpMtSessionTable 1 }

MplsLdpMtSessionEntry ::=
    SEQUENCE {
        mplsLdpMtSessionPeerId
        MplsLdpIdentifier,

```



```

        mplsLdpMtSessionState
            INTEGER,
        mplsLdpMtSessionStateLastChange
            TimeStamp
    }

mplsLdpMtSessionPeerId OBJECT-TYPE
    SYNTAX MplsLdpIdentifier
    MAX-ACCESS not-accessible
    STATUS current
    DESCRIPTION
        "The LDP identifier of this LDP MT Peer."
    ::= { mplsLdpMtSessionEntry 1 }

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

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

mplsLdpMtLspObjects OBJECT IDENTIFIER ::= { mplsLdpMtObjects 3 }

-- mplsLdpMtInSegmentTable
mplsLdpMtInSegmentTable OBJECT-TYPE
    SYNTAX SEQUENCE OF MplsLdpMtInSegmentEntry
    MAX-ACCESS not-accessible
    STATUS current
    DESCRIPTION

```

"This table contains information about the MPLS L
 Distribution Protocol Multi Topology In-Segments
 exist on this Label Switching Router (LSR) or Lab
 Edge Router (LER)."
 ::= { mplsLdpMtLspObjects 1 }

mplsLdpMtInSegmentEntry OBJECT-TYPE
 SYNTAX MplsLdpMtInSegmentEntry
 MAX-ACCESS not-accessible
 STATUS current
 DESCRIPTION
 "An entry in this table represents information on
 LDP MT LSP which is represented by a MT session's
 combination (mplsLdpMtEntityLdpId, mplsLdpMtEntit
 yMtId, mplsLdpMtEntityIndex, mplsLdpMtSessionPeerId).
 The information contained in a row is read-only."
 INDEX { mplsLdpMtEntityLdpId, mplsLdpMtEntityMtId, mplsLd
 pMtEntityIndex,
 mplsLdpMtSessionPeerId }
 ::= { mplsLdpMtInSegmentTable 1 }

MplsLdpMtInSegmentEntry ::=
 SEQUENCE {
 mplsLdpMtInSegmentIndex
 MplsIndexType,
 mplsLdpMtInSegmentLabelType
 MplsLdpLabelType,
 mplsLdpMtInSegmentLspType
 MplsLspType
 }

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

mplsLdpMtInSegmentLabelType OBJECT-TYPE
 SYNTAX MplsLdpLabelType
 MAX-ACCESS read-only
 STATUS current
 DESCRIPTION


```

        "The Layer 2 Label Type."
 ::= { mplsLdpMtInSegmentEntry 2 }

mplsLdpMtInSegmentLspType OBJECT-TYPE
    SYNTAX MplsLspType
    MAX-ACCESS read-only
    STATUS current
    DESCRIPTION
        "The type of LSP connection."
 ::= { mplsLdpMtInSegmentEntry 3 }

-- mplsLdpMtInSegmentStatsTable
    mplsLdpMtInSegmentStatsTable OBJECT-TYPE
        SYNTAX SEQUENCE OF MplsLdpMtInSegmentStatsEntry
        MAX-ACCESS not-accessible
        STATUS current
        DESCRIPTION
            "This table contains statistical information for
LDP MT
            in-segments to an LSR."
 ::= { mplsLdpMtLspObjects 2 }

mplsLdpMtInSegmentStatsEntry OBJECT-TYPE
    SYNTAX MplsLdpMtInSegmentStatsEntry
    MAX-ACCESS not-accessible
    STATUS current
    DESCRIPTION
        "An entry in this table contains statistical info
rmation
        about one incoming MT segment which is configured
        in the
        mplsLdpMtInSegmentTable. The counters in this ent
ry should
        behave in a manner similar to that of the MT inte
rface.
        mplsLdpMtInSegmentStatsDiscontinuityTime indicate
s the time
        of the last discontinuity in all of these objects
        ."
    AUGMENTS { mplsLdpMtInSegmentEntry }
 ::= { mplsLdpMtInSegmentStatsTable 1 }

MplsLdpMtInSegmentStatsEntry ::=
    SEQUENCE {
        mplsLdpMtInSegmentStatsOctets
            Counter32,
        mplsLdpMtInSegmentStatsPackets
            Counter32,
        mplsLdpMtInSegmentStatsErrors
            Counter32,
        mplsLdpMtInSegmentStatsDiscards
    }

```



```

        Counter32,
        mplsLdpMtInSegmentStatsHCOctets
        Counter64,
        mplsLdpMtInSegmentStatsDiscontinuityTime
        TimeTicks
    }

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

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

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

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

```



```
::= { mplsLdpMtInSegmentStatsEntry 4 }
```

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

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

```
-- mplsLdpMtOutSegmentTable
mplsLdpMtOutSegmentTable OBJECT-TYPE
    SYNTAX SEQUENCE OF MplsLdpMtOutSegmentEntry
    MAX-ACCESS not-accessible
    STATUS current
    DESCRIPTION
        "This table contains information about the MPLS L
        abel Distribution
        Protocol Multi Topology Out-Segments which exist
        on this Label
        Switching Router (LSR) or Label Edge Router (LER)
        ."
    ::= { mplsLdpMtLspObjects 3 }
```

```
mplsLdpMtOutSegmentEntry OBJECT-TYPE
    SYNTAX MplsLdpMtOutSegmentEntry
    MAX-ACCESS not-accessible
    STATUS current
    DESCRIPTION
        "An entry in this table represents information on
        a single LDP MT
        LSP which is represented by a MT session's index
        combination
```

dpMtEntityIndex, (mplsLdpMtEntityLdpId, mplsLdpMtEntityMtId, mplsL
mplsLdpMtSessionPeerId).

Li, et al.

Expires December 29, 2012

[Page 19]

```

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

MplsLdpMtOutSegmentEntry ::=
    SEQUENCE {
        mplsLdpMtOutSegmentIndex
            MplsIndexType,
        mplsLdpMtOutSegmentLabelType
            MplsLdpLabelType,
        mplsLdpMtOutSegmentLspType
            MplsLspType
    }

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

mplsLdpMtOutSegmentLabelType OBJECT-TYPE
    SYNTAX MplsLdpLabelType
    MAX-ACCESS read-only
    STATUS current
    DESCRIPTION
        "The Layer 2 Label Type."
    ::= { mplsLdpMtOutSegmentEntry 2 }

mplsLdpMtOutSegmentLspType OBJECT-TYPE
    SYNTAX MplsLspType
    MAX-ACCESS read-only
    STATUS current
    DESCRIPTION
        "The type of LSP connection."
    ::= { mplsLdpMtOutSegmentEntry 3 }

-- mplsLdpMtOutSegmentStatsTable
mplsLdpMtOutSegmentStatsTable OBJECT-TYPE
    SYNTAX SEQUENCE OF MplsLdpMtOutSegmentStatsEntry
    MAX-ACCESS not-accessible

```

```

STATUS current
DESCRIPTION
    "This table contains statistical information for
    LDP MT out-segments to an LSR."
::= { mplsLdpMtLspObjects 4 }

```

```

mplsLdpMtOutSegmentStatsEntry OBJECT-TYPE
SYNTAX MplsLdpMtOutSegmentStatsEntry
MAX-ACCESS not-accessible
STATUS current
DESCRIPTION
    "An entry in this table contains statistical info
    rmation
    about one incoming MT segment which is configured
    in the
    mplsLdpMtOutSegmentTable. The counters in this en
    try
    should behave in a manner similar to that of the
    MT interface.
    mplsLdpMtOutSegmentStatsDiscontinuityTime indicat
    es the time
    of the last discontinuity in all of these objects
    ."
AUGMENTS { mplsLdpMtOutSegmentEntry }
::= { mplsLdpMtOutSegmentStatsTable 1 }

```

```

MplsLdpMtOutSegmentStatsEntry ::=
SEQUENCE {
    mplsLdpMtOutSegmentStatsOctets
        Counter32,
    mplsLdpMtOutSegmentStatsPackets
        Counter32,
    mplsLdpMtOutSegmentStatsErrors
        Counter32,
    mplsLdpMtOutSegmentStatsDiscards
        Counter32,
    mplsLdpMtOutSegmentStatsHCOctets
        Counter64,
    mplsLdpMtOutSegmentStatsDiscontinuityTime
        TimeTicks
}

```

```

mplsLdpMtOutSegmentStatsOctets OBJECT-TYPE
SYNTAX Counter32
MAX-ACCESS read-only
STATUS current
DESCRIPTION
    "This value represents the total number of octets
    received by
    this MT segment. It MUST be equal to the least si
    gnificant 32 bits
    of mplsLdpMtOutSegmentStatsHCOctets if mplsLdpMtO
    utSegmentStatsHCOctets
    is supported according to the rules spelled out i
    n RFC2863."
::= { mplsLdpMtOutSegmentStatsEntry 1 }

```



```

mplsLdpMtOutSegmentStatsPackets OBJECT-TYPE
    SYNTAX Counter32
    MAX-ACCESS read-only
    STATUS current
    DESCRIPTION
        "Total number of packets received by this MT segm
ent."
    ::= { mplsLdpMtOutSegmentStatsEntry 2 }

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

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

mplsLdpMtOutSegmentStatsHCOctets OBJECT-TYPE
    SYNTAX Counter64
    MAX-ACCESS read-only
    STATUS current
    DESCRIPTION
        "The total number of octets received. This is the
        64 bit version
        of mplsLdpMtOutSegmentStatsOctets, if mplsLdpMtOu
tSegmentStatsHCOctets
        is supported according to the rules spelled out i
n RFC2863."
    ::= { mplsLdpMtOutSegmentStatsEntry 5 }

mplsLdpMtOutSegmentStatsDiscontinuityTime OBJECT-TYPE
    SYNTAX TimeTicks
    MAX-ACCESS read-only
    STATUS current
    DESCRIPTION
        "The value of sysUpTime on the most recent occasi
on at which any

```


one or more of this MT segment's Counter32 or Counter64 suffered a discontinuity. If no such discontinuities have occurred since the last re-initialization of the local management subsystem, then this object contains a zero value."

```
 ::= { mplsLdpMtOutSegmentStatsEntry 6 }
```

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

        If no such changes have occurred since the last re-initialization of the local management subsystem, then this object contains a zero value."
    ::= { mplsLdpMtLspObjects 5 }
```

```
mplsLdpMtLspIndexNext OBJECT-TYPE
    SYNTAX IndexIntegerNextFree
    MAX-ACCESS read-only
    STATUS current
    DESCRIPTION
        "This object contains an appropriate value to be used for mplsLdpMtLspIndex when creating entries in the mplsLdpMtLspTable.

        The value 0 indicates that no unassigned entries are available."
    ::= { mplsLdpMtLspObjects 6 }
```

```
-- mplsLdpMtLspTable
mplsLdpMtLspTable OBJECT-TYPE
    SYNTAX SEQUENCE OF MplsLdpMtLspEntry
    MAX-ACCESS not-accessible
    STATUS current
    DESCRIPTION
        "This table specifies MT LIB label switching information. Entries in this table define LIB label switching entries associated with the specified topology."
    ::= { mplsLdpMtLspObjects 7 }
```

```
mplsLdpMtLspEntry OBJECT-TYPE
```

SYNTAX MplsLdpMtLspEntry
MAX-ACCESS not-accessible
STATUS current

Li, et al.

Expires December 29, 2012

[Page 23]

```

DESCRIPTION
    "An entry in this table is created by an LSR for
every label within
    the context of a specific topology capable of sup
porting MT LDP LSP.
    The indexing provides an ordering of topologies p
er interface."
INDEX { mplsLdpMtEntityLdpId, mplsLdpMtEntityMtId, mplsLd
pMtEntityIndex,
        mplsLdpMtLspInSegmentIndex, mplsLdpMtLspOutSegmen
tIndex,
        mplsLdpMtLspIndex }
::= { mplsLdpMtLspTable 1 }

MplsLdpMtLspEntry ::=
    SEQUENCE {
        mplsLdpMtLspIndex
            IndexInteger,
        mplsLdpMtLspFecAddr
            InetAddress,
        mplsLdpMtLspFecAddrLength
            InetAddressPrefixLength,
        mplsLdpMtLspInSegmentIndex
            MplsIndexType,
        mplsLdpMtLspOutSegmentIndex
            MplsIndexType,
        mplsLdpMtLspRowStatus
            Integer32,
        mplsLdpMtLspStorageType
            StorageType,
        mplsLdpMtLspOperStatus
            RowStatus
    }

mplsLdpMtLspIndex OBJECT-TYPE
    SYNTAX IndexInteger
    MAX-ACCESS not-accessible
    STATUS current
    DESCRIPTION
        "The index which uniquely identifies this entry."
    ::= { mplsLdpMtLspEntry 1 }

mplsLdpMtLspFecAddr OBJECT-TYPE
    SYNTAX InetAddress
    MAX-ACCESS read-create
    STATUS current
    DESCRIPTION
        "The FEC address of this LDP MT LSP. Note that th
e
        value of this object is interpreted as prefix add
ress."
    REFERENCE
        "RFC 5036, Section 3.4.1 FEC TLV."

```



```
::= { mplsLdpMtLspEntry 2 }
```

```
mplsLdpMtLspFecAddrLength OBJECT-TYPE
    SYNTAX InetAddressPrefixLength
    MAX-ACCESS read-create
    STATUS current
    DESCRIPTION
        "The FEC prefix length of this LDP MT LSP."
    REFERENCE
        "RFC5036, Section 3.4.1. FEC TLV"
    ::= { mplsLdpMtLspEntry 3 }
```

```
mplsLdpMtLspInSegmentIndex OBJECT-TYPE
    SYNTAX MplsIndexType
    MAX-ACCESS not-accessible
    STATUS current
    DESCRIPTION
        "Index of in-segment for this LDP MT LSP."
    ::= { mplsLdpMtLspEntry 4 }
```

```
mplsLdpMtLspOutSegmentIndex OBJECT-TYPE
    SYNTAX MplsIndexType
    MAX-ACCESS not-accessible
    STATUS current
    DESCRIPTION
        "Index of out-segment for this LDP MT LSP."
    ::= { mplsLdpMtLspEntry 5 }
```

```
mplsLdpMtLspRowStatus OBJECT-TYPE
    SYNTAX Integer32
    MAX-ACCESS read-create
    STATUS current
    DESCRIPTION
        "For creating, modifying, and deleting this row.
        When a row in this table has a row in the active(
1)
        state, no objects in this row except this object
        and the mplsLdpMtLspStorageType can be modified."
    ::= { mplsLdpMtLspEntry 6 }
```

```
mplsLdpMtLspStorageType OBJECT-TYPE
    SYNTAX StorageType
    MAX-ACCESS read-create
    STATUS current
```

DESCRIPTION

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

DEFVAL { nonVolatile }
 ::= { mplsLdpMtLspEntry 7 }

mplsLdpMtLspOperStatus OBJECT-TYPE

SYNTAX RowStatus
 MAX-ACCESS read-create
 STATUS current

DESCRIPTION

"The status of this conceptual row. If the value of this object is 'active(1)', then none of the writable objects of this entry can be modified, except to set this object to 'destroy(6)'."

NOTE: if this row is being referenced by any entry in the mplsLdpLspFecTable, then a request to destroy this row, will result in an inconsistentValue error."

::= { mplsLdpMtLspEntry 8 }

mplsLdpMtConformance OBJECT IDENTIFIER ::= { mplsLdpMtStdMIB 2 }

mplsLdpMtGroups OBJECT IDENTIFIER ::= { mplsLdpMtConformance 1 }

mplsLdpMtEntityGroup OBJECT-GROUP

OBJECTS { mplsLdpMtEntityLastChange, mplsLdpMtEntityIndexNext, mplsLdpMtEntityMtId, mplsLdpMtEntityAdminStatus, mplsLdpMtEntityStorageType, mplsLdpMtEntityRowStatus, mplsLdpMtEntityStatsDiscontinuityTime, mplsLdpMtEntityStatsHCOctets, mplsLdpMtEntityStatsDiscards, mplsLdpMtEntityStatsErrors, mplsLdpMtEntityStatsPackets, mplsLdpMtEntityStatsOctets }

STATUS current

DESCRIPTION

"Objects that apply to all MPLS LDP MT Entity implementations."

::= { mplsLdpMtGroups 2 }

mplsLdpMtSessionGroup OBJECT-GROUP

OBJECTS { mplsLdpMtSessionLastChange, mplsLdpMtSessionState, mplsLdpMtSessionStateLastChange }

STATUS current
DESCRIPTION

Li, et al.

Expires December 29, 2012

[Page 26]

"Objects that apply to all MPLS LDP MT Session implementations."

::= { mplsLdpMtGroups 3 }

```

mplsLdpMtLspGroup OBJECT-GROUP
    OBJECTS { mplsLdpMtLspLastChange, mplsLdpMtLspIndexNext,
mplsLdpMtLspFecAddr,
    mplsLdpMtLspFecAddrLength, mplsLdpMtLspRowStatus,
    mplsLdpMtLspStorageType, mplsLdpMtLspOperStatus,
    mplsLdpMtInSegmentIndex, mplsLdpMtInSegmentLabelType,
    mplsLdpMtInSegmentLspType, mplsLdpMtInSegmentStatsOctets,
    mplsLdpMtInSegmentStatsPackets, mplsLdpMtInSegmentStatsErrors,
    mplsLdpMtInSegmentStatsDiscards, mplsLdpMtInSegmentStatsHCOctets,
    mplsLdpMtInSegmentStatsDiscontinuityTime, mplsLdpMtOutSegmentIndex,
    mplsLdpMtOutSegmentLabelType, mplsLdpMtOutSegmentLspType,
    mplsLdpMtOutSegmentStatsOctets, mplsLdpMtOutSegmentStatsPackets,
    mplsLdpMtOutSegmentStatsErrors, mplsLdpMtOutSegmentStatsDiscards,
    mplsLdpMtOutSegmentStatsHCOctets, mplsLdpMtOutSegmentStatsDiscontinuityTime
    }

```

STATUS current

DESCRIPTION

"Objects that apply to all MPLS LDP MT LSP implementations."

::= { mplsLdpMtGroups 4 }

```

mplsLdpMtNotificationGroup NOTIFICATION-GROUP
    NOTIFICATIONS { mplsLdpMtLspUp, mplsLdpMtLspDown }
    STATUS current
    DESCRIPTION

```

"The notifications for an MPLS LDP MT implementation."

::= { mplsLdpMtGroups 5 }

mplsLdpMtCompliances OBJECT IDENTIFIER ::= { mplsLdpMtConformance 2 }

mplsLdpMtModuleFullCompliance MODULE-COMPLIANCE

STATUS current

DESCRIPTION

"The Module is implemented with support for read-create and read-write. In other words, both monitoring and configuration are available when using this MODULE-COMPLIANCE."

```

tSessionGroup,
tificationGroup }

MODULE -- this module
    MANDATORY-GROUPS { mplsLdpMtEntityGroup, mplsLdpM
                        mplsLdpMtLspGroup, mplsLdpMtNo
::= { mplsLdpMtCompliances 1 }

```

```
mplsLdpMtModuleReadOnlyCompliance MODULE-COMPLIANCE
    STATUS current
    DESCRIPTION
        "The Module is implemented with support for read-
only.
        In other words, only monitoring is available by i
mplementing
        this MODULE-COMPLIANCE"
    MODULE -- this module
        MANDATORY-GROUPS { mplsLdpMtEntityGroup, mplsLdpM
tSessionGroup,
                                mplsLdpMtLspGroup, mplsLdpMtNo
tificationGroup }
    ::= { mplsLdpMtCompliances 2 }
```

END

5. Security Considerations

It needs to be further identified.

6. IANA Considerations

There is no necessary to request new IANA code in the draft.

7. Normative References

- [RFC3813] Srinivasan, C., Viswanathan, A., and T. Nadeau, "Multiprotocol Label Switching (MPLS) Label Switching Router (LSR) Management Information Base (MIB)", RFC 3813, June 2004.
- [RFC3814] Nadeau, T., Srinivasan, C., and A. Viswanathan, "Multiprotocol Label Switching (MPLS) Forwarding Equivalence Class To Next Hop Label Forwarding Entry (FEC-To-NHLFE) Management Information Base (MIB)", RFC 3814, June 2004.
- [RFC3815] Cucchiara, J., Sjostrand, H., and J. Luciani, "Definitions of Managed Objects for the Multiprotocol Label Switching (MPLS), Label Distribution Protocol (LDP)", RFC 3815, June 2004.
- [RFC5036] Andersson, L., Minei, I., and B. Thomas, "LDP Specification", RFC 5036, October 2007.

[RFC3410] Case, J., Mundy, R., Partain, D., and B. Stewart,
"Introduction and Applicability Statements for Internet-
Standard Management Framework", RFC 3410, December 2002.

[I-D.ietf-mpls-ldp-multi-topology]
Zhao, Q., Fang, L., Zhou, C., Li, L., and N. So, "LDP
Extensions for Multi Topology Routing",
draft-ietf-mpls-ldp-multi-topology-03 (work in progress),
March 2012.

Authors' Addresses

Chen Li
China Mobile
Unit2, Dacheng Plaza, No. 28 Xuanwumenxi Ave, Xuanwu District
Beijing 100053
P.R. China

Email: lichenyj@chinamobile.com

Lianyuan Li
China Mobile
Unit2, Dacheng Plaza, No. 28 Xuanwumenxi Ave, Xuanwu District
Beijing 100053
P.R. China

Email: lilianyuan@chinamobile.com

Lu Huang
China Mobile
Unit2, Dacheng Plaza, No. 28 Xuanwumenxi Ave, Xuanwu District
Xunwu District, Beijing 100053
China

Email: huanglu@chinamobile.com

Emily Chen
Huawei Technologies
2330 Central Expressway
Santa Clara, CA 95050
US

Email: emily.chenying@huawei.com

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

Email: quintin.zhao@huawei.com

MPLS Working Group
Internet-Draft
Intended status: Standards Track
Expires: January 8, 2013

R. Li
Q. Zhao
Huawei Technologies
C. Jacquenet
France Telecom Orange
July 7, 2012

Receiver-Driven Multicast Traffic-Engineered Label-Switched Paths
draft-lzj-mpls-receiver-driven-multicast-rsvp-te-01.txt

Abstract

This document describes extensions to Resource Reservation Protocol - Traffic Engineering (RSVP-TE) for the setup of Receiver-Driven Traffic-Engineered point-to-multipoint (P2MP) and multipoint-to-multipoint (MP2MP) Label Switched Paths (LSPs) in Multi-Protocol Label Switching (MPLS) and Generalized MPLS (GMPLS) networks.

Status of this Memo

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

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

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

This Internet-Draft will expire on January 8, 2013.

Copyright Notice

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

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of

the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

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

Table of Contents

1. Introduction	4
1.1. Motivation	4
1.2. Terminology	4
1.3. Overview	5
2. Receiver-Driven mRSVP-TE LSP Examples	7
2.1. P2MP Example	8
2.2. MP2MP Example	9
3. Signaling Protocol Extensions	10
3.1. Mechanisms	10
3.1.1. Sessions	10
3.1.2. L2S Sub-LSPs	11
3.1.3. Path Originator and Data Receiver	12
3.1.4. Explicit Routing	13
3.2. Path Messages	13
3.3. Resv Messages	15
3.4. PathErr Messages	15
3.5. ResvErr Message	15
3.6. PathTear Messages	16
4. New and Updated Objects	16
4.1. SESSION Objects	16
4.1.1. P2MP LSP for IPv4 SESSION Objects	16
4.1.2. MP2MP LSP for IPv4 SESSION Objects	17
4.1.3. P2MP LSP for IPv6 SESSION Objects	17
4.1.4. MP2MP LSP for IPv6 SESSION Objects	17
4.2. SENDER_TEMPLATE Objects	18
4.2.1. Multicast LSP IPv4 SENDER_TEMPLATE Objects	18
4.2.2. Multicast LSP IPv6 SENDER_TEMPLATE Objects	18
4.3. L2S_SUB_LSP Objects	19
4.3.1. L2S_SUB_LSP IPv4 Objects	19
4.3.2. L2S_SUB_LSP IPv6 Objects	19
4.4. FILTER_SPEC Objects	20
4.4.1. mRSVP-TE LSP_IPv4 FILTER_SPEC Objects	20
4.4.2. mRSVP-TE LSP_IPv6 FILTER_SPEC Objects	20
5. Applications	20
5.1. Interwork with PIM	20
5.2. Multicast VPN	21
6. Fast Re-Route Considerations	21
7. Backward Compatibility	21
8. Acknowledgements	22
9. IANA Considerations	22
10. Security Considerations	22
11. References	23
11.1. Normative References	23
11.2. Informative References	24
Authors' Addresses	24

1. Introduction

Multiparty multimedia applications are getting greater attention in the telecom and datacom world. Such applications are QoS-demanding and can therefore benefit from the activation of MPLS traffic engineering capabilities that lead to the dynamic computation and establishment of MPLS LSPs whose characteristics comply with application-specific QoS requirements. P2MP-TE [RFC4875] defines a procedure to set up point-to-multipoint LSPs from sender to receivers. Sometimes multicast data streams are required to get transported over both IP networks and MPLS networks, which require PIM must interwork with RSVP-TE. On other times, PIM bootstrapping messages need to transport over an intermediate MPLS domain. This document extends RSVP-TE for the dynamic computation of receiver-driven P2MP and MP2MP LSP tree structures.

1.1. Motivation

IP multicast distribution trees are receiver-initiated and dynamic by nature. IP multicast-enabled applications are also bandwidth savvy, especially in the area of residential IPTV services, where the delivery of multicast contents to several hundreds of thousands of IPTV receivers assumes the appropriate level of quality.

Current source-driven P2MP LSP establishment, as defined as in [RFC4875], assumes a priori knowledge of receiver locations, and the LSP signalling is initiated and driven by the data sender(headend). The priori knowledge of receiver locations is obtained either through static configuration or by using another protocol to discover such receivers. On the other hand, there is no straightforward way to support MP2MP applications by using P2MP LSP unless full-meshed P2MP LSPs are set up.

The receiver-driven extension to RSVP-TE defined in this document will support both P2MP LSPs and MP2MP LSPs. Moreover, it does not require the sender to know all the receivers' locations a priori. The protocols for discovery of receivers are not needed. It provides a natural mechanism to interwork with PIM dynamically.

1.2. Terminology

The following terms are used in this document:

- o Sender: Sender refers to the Originator (and hence the Sender) of the content/payload, as defined in [RFC2205].
- o Receiver: Receiver refers to the Receiver of the content/payload, as defined in [RFC2205].

- o Upstream: The direction of flow from content Receiver toward content Sender, as defined in [RFC2205].
- o Downstream: The direction of flow from content Sender toward content Receiver, as defined in [RFC2205].
- o Path-Sender: The sender of RSVP PATH messages, with no correlation to the direction of content/payload flows. Its flow direction is irrelevant to that of Sender defined above. All other control messages discussed in this document will use this as the reference.
- o Path-Receiver: The receiver of RSVP PATH messages, with no correlation to the direction of content/payload flows.
- o Path-Initiator: The Path-Sender that originated a RSVP PATH message. This is different from Path-Sender in that an intermediate node can be a Path-Sender, but such an intermediate node cannot create and initiate the RSVP PATH message. A Path-Initiator is a Path-Sender, but a Path-Sender doesn't have to be a Path-Initiator.
- o Path-Terminator: The Path-Receiver that does NOT propagate the Path message any further. This is different from Path-Receiver in that an intermediate node can be a Path-Receiver, but such an intermediate node will propagate the Path message to the next hop.
- o Root: A router where a multicast LSP tree is rooted at. Data enters the root and then is distributed to leaves along the P2MP/MP2MP LSP.

1.3. Overview

Although the receiver-driven extensions to RSVP-TE as defined in this document use the existing sender-driven syntax, there are important semantic differences that need to be defined for correct interpretation and interoperability. In the receiver-driven context, we inverted the semantics of RSVP-TE messages, while keeping the syntax unchanged as much as possible. We will use mRSVP-TE to represent the RSVP-TE with receiver-driven extensions described in this document.

The following are some key differences that are specific to the receiver-driven paradigm:

- o The leaf router: the router that receives data/content/payload. In this document, the leaf router will initiate PATH messages. In some sense, the leaf router and the receiver mean the same thing.

The term "receiver-driven" also means "leaf-driven".

- o L2S Destinations: routers where user data payload traffic enters the LSP. L2S means Leaf-to-Source. The source is the sender or root of a multicast stream.
- o RSVP P2MP PATH messages traverse from receivers to the root.
- o RSVP P2MP RESV messages traverse from the root to the leaf routers of the P2MP tree structure.
- o A RSVP RESV message received by a router is interpreted as a successful resource reservation made by the upstream node for the establishment of the P2MP tree structure.
- o A RSVP RESV message received by a router is interpreted as successful resource reservation made by the downstream node for the establishment of an MP2MP tree structure.
- o Label allocation on incoming interfaces is done prior to sending RSVP PATH messages upstream for P2MP tree structures.
- o Label allocation on incoming interfaces is done prior to sending RSVP RESV messages upstream for MP2MP tree structures.
- o For P2MP LSP tree structures, a node receiving a RSVP PATH message first decides if this RSVP PATH message will make the said node a branch LSR or not. If it is not a branch LSR, it is a transit LSR. In the case that it will become a transit LSR because of this PATH message, it will, before sending the RSVP PATH message upstream, allocate required bandwidth on the interface on which the RSVP PATH message is received. The upstream node can send traffic soon after successfully reserving resources on the downstream link, on which the RSVP PATH message SHOULD be received. In the case that the node is already a branch or a transit node before it receives the PATH message, then it will allocate required bandwidth on the interface on which the RSVP PATH message is received, and send the RESV message to the node which sends the PATH message without propagating the PATH message further to the upstream node. For P2MP LSPs, a label is carried by the PATH message and should be used by the upstream node when distributing the data from upstream to downstream.
- o For MP2MP LSP tree structures, a node will allocate required bandwidth on the interface through which the RSVP PATH message is sent before sending the RSVP PATH message upstream. A node receiving a RSVP PATH message MUST first decide if this RSVP PATH message will make the said node a branch LSR or not. In the case

it will become a transit LSR because of this PATH message, then it will allocate required bandwidth on the interface on which the RSVP PATH message is received and will allocate required bandwidth on the interface through which the RSVP PATH message is sent, before sending the RSVP PATH message upstream. The downstream node can send traffic soon after successfully reserving bandwidth on the upstream link through which the RSVP PATH message SHOULD be sent. The upstream node can send traffic soon after successfully reserving bandwidth on the downstream link on which the RSVP PATH message SHOULD be received. In the case that the node is already a branch or a transit node before it receives the PATH message, then it will allocate required resources on the interface on which the RSVP PATH message is received, and send the RESV message to the node which sends the PATH message without propagating the PATH message further to the upstream node. The label carried by the PATH message should be used by the Path-Receiver node to forward data from the Path-Receiver node to the Path-Sender node, and the label carried by RESV messages should be used by its corresponding Path-Sender node to send data from the Path-Sender node to the Path-Receiver node.

- o For the sake of readability, from now on all mRSVP-TE LSPs will be used to represent all P2MP and/or MP2MP LSPs in receiver-driven (RD) multicast P2MP/MP2MP MPLS environments. We will sometimes use RD P2MP TE LSP or RD MP2MP TE LSP to represent such receiver-driven multicast LSPs.

2. Receiver-Driven mRSVP-TE LSP Examples

In what follows we describe two examples to show how P2MP and MP2MP are set up, respectively. In both of such examples, Path messages are initiated by data receivers.

For the P2MP example, a Path message carries a label for the use of sending data downstream. And for the MP2MP example, both Path message and Resv message carries a label for sending data downstream and upstream.

2.1. P2MP Example

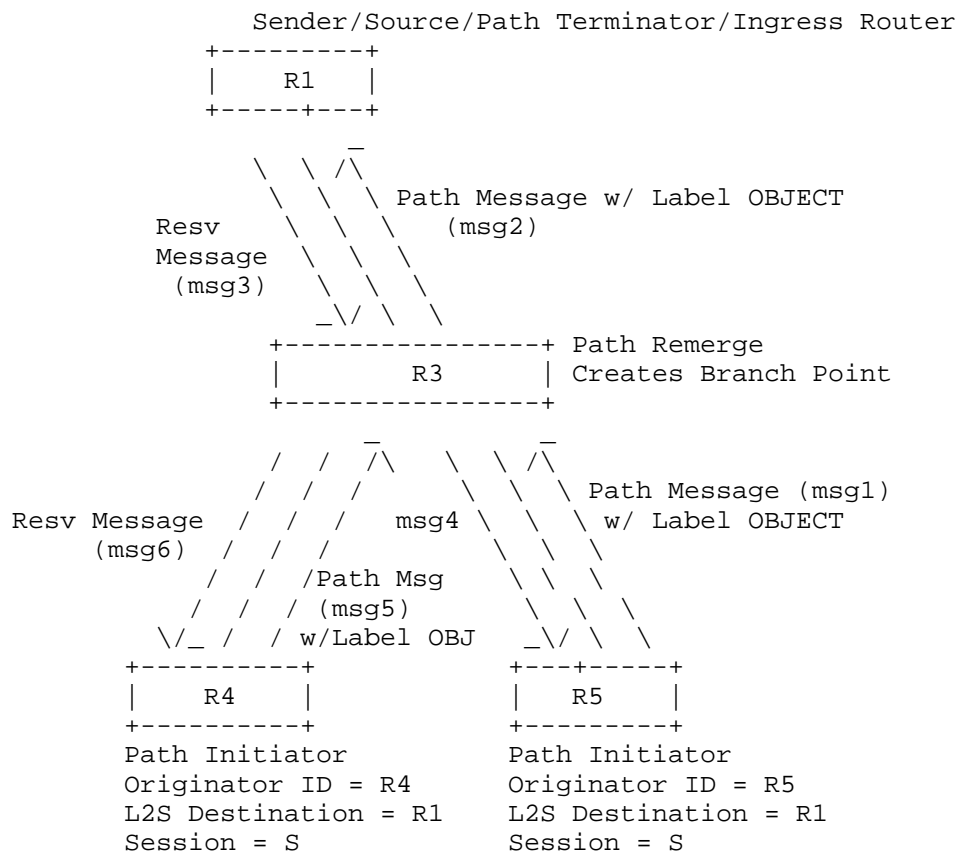


Figure 1: P2MP Example

In Figure 1, when R5 is added as the first leaf of a mulitcast distribution tree (multicast LSP), the message flow goes as follows: R5->msg1->R3->msg2->R1->msg3->R3->msg4->R5. When the leaf R4 is added, the message flow goes from R4->msg5->R3->msg6->R4. In this case, when R3 receives msg5, R3 finds out that a multicast LSP has already been set up for the same session and the same source. Therefore, R3 finds itself a branch node for leaf R4 and R5, so it will terminate the PATH message and build the corresponding RESV message and send it back to R4. The association of the LSP initiated by R4 to the existing multicast LSP is determined based on the

processing of the SESSION object and L2S_SUB_LSP object from the mRSVP-TE message. The SESSION object and the L2S_SUB_LSP objects are documented later in this draft.

2.2. MP2MP Example

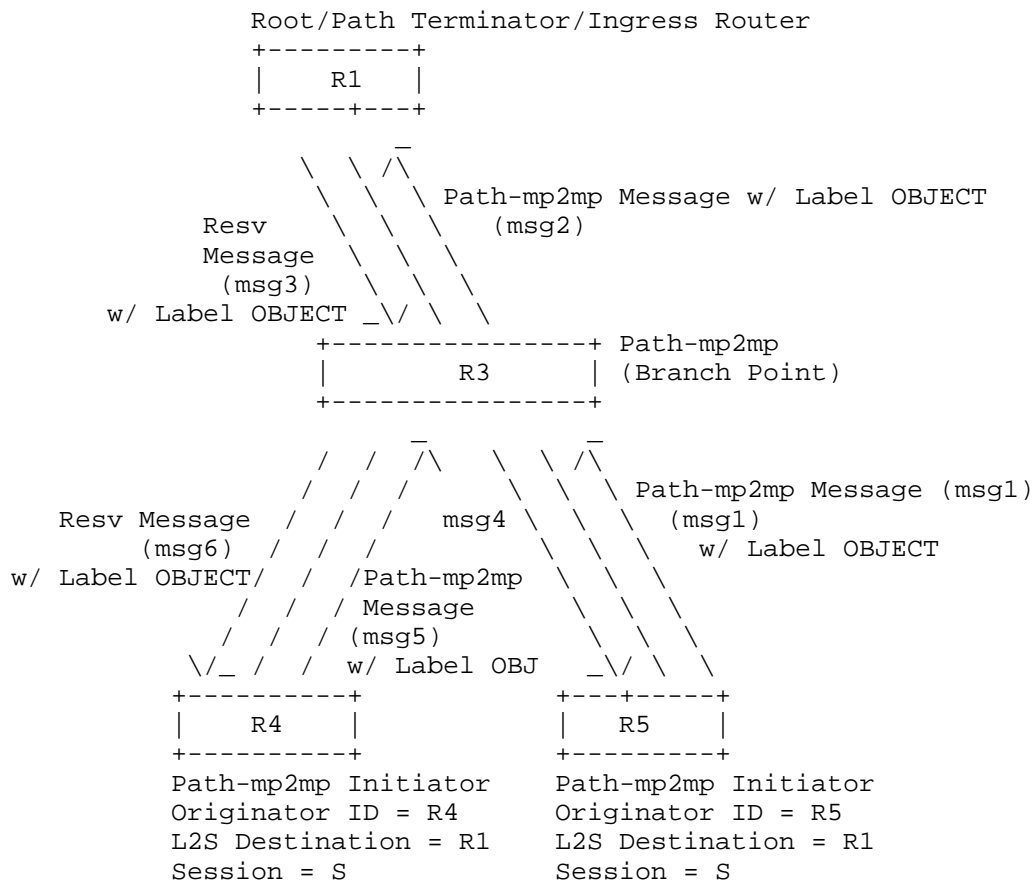


Figure 2: MP2MP Example

In Figure 2, when R5 is added as the first leaf (as both a sender and a receiver) of an MP2MP multicast LSP, the message flow goes from R5->msg1->R3->msg2->R1->msg3->R3->msg4->R5. When the leaf R4 (as both a sender and a receiver) is added, the message flow goes from R4->msg5->R3->msg6->R4. In this case, when R3 receives msg5, R3 finds out that an MP2MP mulitcast LSP has already been set up for the same session and the same root and R3 will become the branch LSR for

the leaf R4 and R5, so it will terminate the PATH message, build a RESV message and send the RESV message back to R4. The association of the LSP initiated by R4 to the existing MP2MP LSP is determined based on the processing of the SESSION object and the S2L_SUB_LSP from the mRSVP-TE message. The SESSION objects and the L2S_SUB_LSP objects are further documented later in this draft.

3. Signaling Protocol Extensions

The RSVP-TE with receiver-driven extensions (mRSVP-TE) is similar to the RSVP-TE protocol as specified in [RFC4875], [RFC3473] and [RFC3209], but differs in that the data receivers of an LSP tunnel initiate the Path messages toward the data sender (or the root of a multicast LSP). Compared with [RFC4875], mRSVP-TE can also be used to set up MP2MP LSPs.

In the context of the receiver-driven RSVP-TE, the Receiver is the Path-Originator. The Path messages go from the Receivers towards the Sender. The Resv messages flow in the opposite direction as compared to the Path messages, i.e. Resv messages are generated by the Sender or a branch LSR. Path messages flow in opposite directions as compared with those of the multicast stream distributions, while Resv messages flow in the same directions as the multicast streams.

In the context of the receiver-driven RSVP-TE, a Path message will be terminated at the "root" of the multicast distribution tree (multicast LSP) or at an intermediate node if the intermediate node has received another Path message from another receiver for the same multicast distribution tree. When an intermediate node receives two or more Path messages for the same multicast distribution tree, the intermediate node will merge them together. Whether two Path messages should be merged depends on the information encoded in the SESSION and L2S-SUB-LSP objects. The SESSION object encodes multicast group information and the L2S-SUB-LSP (leaf-to-source sub-lsp) object encodes the multicast source or multicast root information.

The following sections describe the receiver-driven extensions to the RSVP-TE protocol. When there is no difference in the protocol, the usage of [RFC4875] is assumed.

3.1. Mechanisms

3.1.1. Sessions

As specified in [RFC2205], a session is a data flow with a particular destination and transport-layer protocol. In the context of

multicast, the data flow is essentially a multicast distribution tree rooted at the P2MP source or MP2MP root.

For the sake of reliability, two or more sources/roots may be deployed to distribute the same multicast streams. A multicast stream is often represented by a multicast group address. In this document, we will encode the multicast group address in the SESSION object and the multicast source/root address in the leaf-to-source sub-LSP object. Note that the same session can have different sources/roots, and the same sources/roots can have different sessions.

In the context of the receiver-driven mRSVP-TE, the processing of SESSION objects is different from that of SESSION objects in sender-driven RSVP-TE [RFC4875]. In order to distinguish them, we will employ different C-Types of SESSIONs. In this document we will document SESSION objects for native IPv4/IPv6 multicast applications. For new and more applications, new types of SESSION objects will be added.

Following the method used by RSVP-TE and P2MP RSVP-TE, this draft documents the use of some new SESSION C-Type as follows:

```
Class Name = SESSION
C-Type
  XX+0    mRSVP_TE_P2MP_LSP_TUNNEL_IPv4 C-Type
  XX+1    mRSVP_TE_P2MP_LSP_TUNNEL_IPv6 C-Type
  XX+2    mRSVP_TE_MP2MP_LSP_TUNNEL_IPv4 C-Type
  XX+3    mRSVP_TE_MP2MP_LSP_TUNNEL_IPv6 C-Type
```

Where XX is a number to be allocated by IANA.

Figure 3: New C-Types of SESSIONs

The new SESSION C-Type MUST be used in all receiver-driven P2MP RSVP-TE messages.

3.1.2. L2S Sub-LSPs

A multicast LSP is composed of one or more leaf-to-source sub-LSPs, which are merged together at the branch nodes. There are two ways to identify each such sub-LSP:

- o From the Sender's perspective, each sub-LSP is identified by the SESSION object, the SENDER_TEMPLATE object and S2L_SUB_LSP object, as specified in [RFC 4875]. The SESSION object encodes P2MP ID,

Tunnel ID, and Extended Tunnel ID. The P2MP ID is unique within the scope of the sender (ingress LSR) and remains constant throughout the lifetime of the P2MP tree structure. The Extended Tunnel ID, which remains constant throughout the lifetime of the P2MP tree structure, and which should contain the sender's address to make sure the identifier is globally unique. Finally, the Tunnel ID, also remains constant throughout the lifetime of the P2MP tree structure. The SENDER_TEMPLATE object contains the ingress LSR source address. The S2L_SUB_LSP contains the destination address of the sub-LSP.

- o From the Receiver's perspective, each sub-LSP is identified by a new SESSION object, a new SENDER_TEMPLATE object and a new L2S_SUB_LSP object. The SESSION object, different from the one used in typical sender-driven environments, contains information to be used as the key to associate different PATH messages originated from different leaves. The SENDER_TEMPLATE object contains the Path-Originator's address, which is actually the Data Receiver. The L2S_SUB_LSP contains the source or root address of the sub-LSP, i.e. the data Sender's address. The SESSION, SENDER_TEMPLATE and L2S_SUB_LSP all together will identify the multicast stream, the multicast stream's source, and a mulitcast stream's receiver

This document takes the approach from the Receiver's perspective. The approach from the Sender's perspective is documented in [RFC 4875].

Once an LSR receives a receiver-driven Path message with the SESSION object and L2S_SUB_LSP object, the LSR should be able to use the SESSION object and L2S_SUB_LSP object to determine whether the sub-LSP signaled by this Path message should be merged with existing multicast LSPs.

3.1.3. Path Originator and Data Receiver

In the context of the receiver-driven RSVP-TE, a Path Originator is also a Data Receiver. This document will document a new type of SENDER_TEMPLATE object, which contains the Path-Originator's IP address and describes the identity of the Path Originator.

In [RFC 2205] and [RFC 4875], the "sender" is both a path originator and a data sender. In the receiver-driven context, path originators and data senders may be different. For P2MP, path originators are actually the data receivers. For MP2MP, path originators are also both the data senders and data receivers.

In this document, we will use the same Object Class SENDER_TEMPLATE

with a different C-Type to represent and identify Path Originator. In the case of P2MP LSP, the SENDER_TEMPLATE describes the identify of a data receiver. In the case of MP2MP, the SENDER_TEMPLATE describes the identify of an LSR which work as both a data sender and a data receiver.

All of the SESSION object, L2S_SUB_LSP object and SENDER_TEMPLATE object together contained in a Path message will uniquely identify a leaf-to-source sub-LSP.

3.1.4. Explicit Routing

An EXPLICIT_ROUTE Object (ERO) is used to optionally specify the explicit route of an L2S sub-LSP. Each signaled ERO corresponds to a particular L2S_SUB_LSP object. Details of explicit route encoding are specified in section 4.5 of [RFC4875], but they are encoded in a reverse order in the receiver-driven context.

When a Path message signals a L2S sub-LSP, the EXPLICIT_ROUTE object encodes the path from the leaf to the root LSR. The Path message also includes the L2S_SUB_LSP object for the L2S sub-LSP being signaled. The < [<EXPLICIT_ROUTE>], <L2S_SUB_LSP>> tuple represents the L2S sub-LSP and is referred to as the sub-LSP descriptor.

The absence of the ERO should be interpreted as requiring hop-by-hop reverse-forwarding for the sub-LSP based on the root address field of the L2S_SUB_LSP object.

3.2. Path Messages

The mechanism specified in this document allows a multicast P2MP/MP2MP LSP to be signaled using one or more Path messages. Each Path message may signal one L2S sub-LSPs.

A receiver-driven P2MP MPLS-TE LSP uses the Path message to carry the LABEL object upstream from the Receiver towards the Sender. With a receiver-driven usage of the RSVP PATH messages, the LABEL_REQUEST object carried by the PATH message is no longer mandatory, it becomes optional for receiver-driven PATH messages, as specified in Figure 4:

```

<Path Message> ::=      <Common Header> [ <INTEGRITY> ]
                        [ [ <MESSAGE_ID_ACK> | <MESSAGE_ID_NACK> ] ... ]
                        [ <MESSAGE_ID> ]
                        <SESSION> <RSVP_HOP>
                        <TIME_VALUES>
                        [ <EXPLICIT_ROUTE> ]
                        [ <LABEL_REQUEST> ]
                        [ <PROTECTION> ]
                        [ <LABEL_SET> ... ]
                        [ <SESSION_ATTRIBUTE> ]
                        [ <NOTIFY_REQUEST> ]
                        [ <ADMIN_STATUS> ]
                        [ <POLICY_DATA> ... ]
                        <sender descriptor>
                        [ <L2S_SUB_LSP> ]

```

Figure 4: Path Message Extensions

The SESSION object encodes information about the being-signalled multicast stream. The SESSION object together with L2S_SUB_LSP will be used as the key to associate different sub-LSPs to the same multicast LSP.

Using [RFC4875] as the base specification, the LABEL object is added to the <sender descriptor> as specified in Figure 5:

```

<sender descriptor> ::= <SENDER_TEMPLATE> <SENDER_TSPEC>
                        [ <ADSPEC> ]
                        [ <RECORD_ROUTE> ]
                        [ <SUGGESTED_LABEL> ]
                        [ <RECOVERY_LABEL> ]
                        <LABEL>

```

Figure 5: Sender Descriptor

The LABEL object is defined in section 4.1 of [RFC3209]

Note that the receiver-driven Path messages convey the LABEL_REQUEST as an optional object. If the Path message signals a P2MP LSP, the LABEL_REQUEST in the Path message is not used. If the Path message signals an MP2MP, the LABEL_REQUEST is needed to ask for labels from its upstream LSR.

3.3. Resv Messages

Receiver-driven P2MP RSVP-TE does not need any change to the basic RESV messages specified in section 6.1 of [RFC4875], as long as the receiver-driven SESSION objects of the new C-Types are used.

For receiver-driven P2MP LSPs, the Path message carries the LABEL object, and thus the Resv message doesn't have to carry the LABEL object anymore. But for MP2MP LSPs, both Path and Resv messages will carry LABEL objects for sending and receiving purposes, respectively. Within the context of MP2MP LSPs, one of the directions is established as per [RFC3209]. Thus, this document is changing the use of the LABEL object in the FF Flow Descriptor and SE Filter Spec from mandatory to optional, as specified in Figure 6:

```
<FF flow descriptor> ::= [ <FLOWSPEC> ] <FILTER_SPEC> [ <LABEL> ]  
                        [ <RECORD_ROUTE> ]  
                        [ <L2S_SUB_LSP> ]  
  
<SE filter spec> ::=    <FILTER_SPEC> [ <LABEL> ] [ <RECORD_ROUTE> ]  
                        [ <L2S_SUB_LSP> ]
```

Figure 6: Resv Message Extensions

3.4. PathErr Messages

The receiver-driven PathErr messages have the same syntax and utilization as the PathErr message described in [RFC4875], with the difference in the <sender descriptor> carried by the PathErr message. The receiver-driven PathErr message will use the <sender descriptor> defined in this document, the same as that carried by the Path messages which the PathErr messages correspond to.

3.5. ResvErr Message

The receiver-driven ResvErr messages have the same syntax and utilization as the ResvErr message described in [RFC4875]. But the ResvErr messages will be processed as per this document, given that the <FF flow descriptor> and the <SE filter spec> can optionally contain the LABEL object instead of mandating the use of the LABEL object. The optional use of the LABEL object is conditioned by the nature of the multicast LSP, either uni-directional (P2MP) or bi-directional (MP2MP).

3.6. PathTear Messages

The receiver-driven PathTear messages have the same syntax and utilization as the PathTear messages described in [RFC4875] except for the <sender descriptor> carried by the PathTear messages. The receiver-driven PathTear messages will use <sender descriptor> defined in this document, the same as that carried by the Path messages which the PathTear messages correspond to.

4. New and Updated Objects

4.1. SESSION Objects

An mRSVP-TE LSP SESSION object is used to represent a multicast stream whose traffic will be carried by the multicast LSP being set up by the mRSVP-TE. The object still uses the existing SESSION C-Num assigned for RSVP-TE, but new C-Types are defined for the new purposes. Different from the values in the existing point-to-point or point-to-multipoint RSVP-TE SESSION object, the new objects defined by the new C-Types will encode "multicasting" information. The new SESSION object will have enough information so that the Path-Receiver can use the SESSION objects together with L2S_SUB_LSP to determine whether or not to associate different Path messages from different leaves to the same P2MP/MP2MP LSP. The combination of the SESSION object, the SENDER_TEMPLATE object and the L2S_SUB_LSP object will uniquely identify a single L2S sub-LSP.

For native IPv4/IPv6 multicast, IPv4/IPv6 (S, G) or (*, G, RP) will be encoded in the SESSION object for P2MP or MP2MP LSPs. In what follows we specify such session objects for IPv4/IPv6 P2MP and MP2MP applications in the context of receiver-driven RSVP-TE. Other SESSION objects in the receiver-driven context are defined in other documents.

4.1.1. P2MP LSP for IPv4 SESSION Objects

Class = SESSION, mRSVP-TE_P2MP_LSP_TUNNEL_IPv4 C-Type = TBD.

```

      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
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
|                               Multicast Group Address                               |
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+

```

Figure 7: P2MP LSP for IPv4 SESSION Objects

4.1.2. MP2MP LSP for IPv4 SESSION Objects

Class = SESSION, mRSVP_TE_MP2MP_LSP_TUNNEL_IPv4 C-Type = TBD.

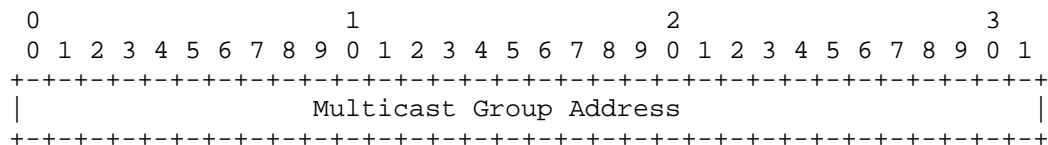


Figure 8: MP2MP LSP for IPv4 SESSION Objects

The MP2MP LSP for IPv4 SESSION objects are of the same format as P2MP LSP for IPv4 SESSION objects, but their C-Types are different.

4.1.3. P2MP LSP for IPv6 SESSION Objects

This is the same as the P2MP LSP for IPv4 SESSION object with the difference that the IPv6 multicast group addresses are 16-byte long.

Class = SESSION, mRSVP_TE_P2MP_LSP_TUNNEL_IPv6 C-Type = TBD.

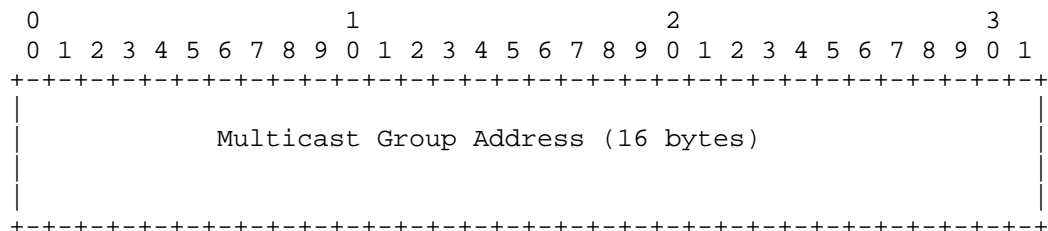


Figure 9: P2MP LSP for IPv6 SESSION Objects

4.1.4. MP2MP LSP for IPv6 SESSION Objects

Class = SESSION, mRSVP_TE_MP2MP_LSP_TUNNEL_IPv6 C-Type = TBD.

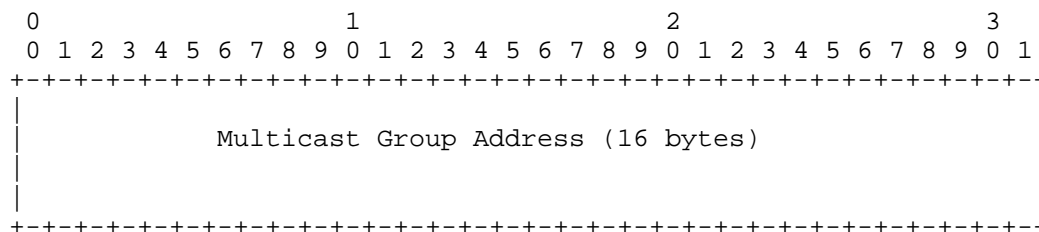


Figure 10: MP2MP LSP for IPv6 SESSION Objects

4.2. SENDER_TEMPLATE Objects

The SENDER_TEMPLATE object contains the Path-Initiator LSR address. In this document, the Path-Initiator is the same as the Leaf Router or Data Receiver. The LSP ID can be changed to allow a sender to do a certain level of resource sharing. Thus, multiple instances of the same multicast LSP can be created, each with a different LSP ID. The instances can share resources with each other. The L2S sub-LSPs corresponding to a particular instance use the same LSP ID.

4.2.1. Multicast LSP IPv4 SENDER_TEMPLATE Objects

Class = SENDER_TEMPLATE, mRSVP-TE_LSP_TUNNEL_IPv4 C-Type = TBD.

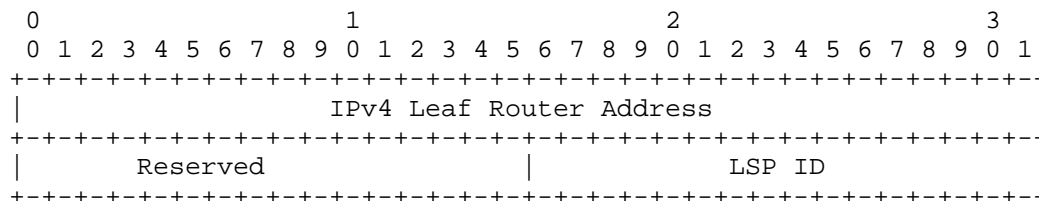


Figure 11: mRSVP-TE Multicast LSP SENDER_TEMPLATE Objects

IPv4 Leaf Router Address: The IPv4 address of the Data Receiver.

LSP ID: A 2-byte identifier that can be changed to allow it to share resources with itself. Its usage is the same as that described in [RFC3209].

4.2.2. Multicast LSP IPv6 SENDER_TEMPLATE Objects

Class = SENDER_TEMPLATE, mRSVP-TE_LSP_TUNNEL_IPv6 C-Type = TBD.

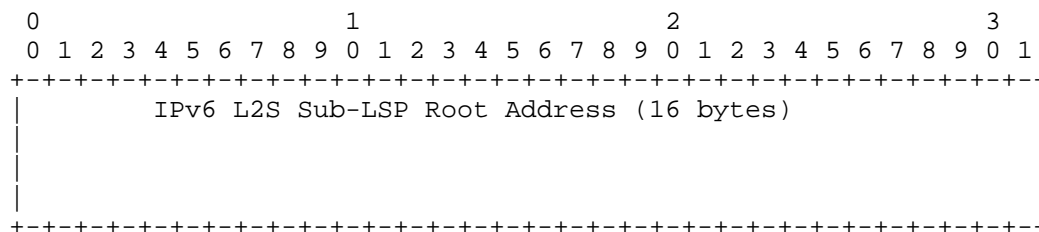


Figure 14: L2S_SUB_LSP IPv6 Object

4.4. FILTER_SPEC Objects

The FILTER_SPEC object is canonical to the SENDER_TEMPLATE object.

4.4.1. mRSVP-TE LSP_IPv4 FILTER_SPEC Objects

Class = FILTER_SPEC, P2MP LSP_IPv4 C-Type = TBD.

The format of the mRSVP-TE LSP_IPv4 FILTER_SPEC object is identical to the mRSVP-TE_LSP_TUNNEL_IPv4 SENDER_TEMPLATE object.

4.4.2. mRSVP-TE LSP_IPv6 FILTER_SPEC Objects

The format of the mRSVP-TE LSP_IPv6 FILTER_SPEC object is identical to the mRSVP-TE_LSP_TUNNEL_IPv6 SENDER_TEMPLATE object.

5. Applications

There are two basic applications for receiver-driven RSVP-TE: interwork with PIM and Multicast VPN.

5.1. Interwork with PIM

Some multicast applications may involve several domains, some of which are operated with PIM while others are enabled with RSVP-TE. This requires the multicast distribution trees to be computed and set up across different domains with PIM and MPLS configured in different domains. When a PIM Join message is received at the border of the MPLS domain, information encoded from the PIM Join message can be encoded as a receiver-driven RSVP-TE Path message which will set up a multicast distribution LSP across the MPLS domain. The root of such a multicast LSP can encode a PIM Join message by using the information encoded in the RSVP-TE Path message. The result of doing so will enable to build a mulitcast distribution tree across both IP and MPLS domains. The multicast tree will consist of a set of IP multicast sub-trees built by PIM and a set of MPLS multicast LSPs

built by the receiver-driven RSVP-TE. In PIM, there is a bootstrapping mechanism about the RP. For bootstrap messages, an MP2MP LSP can be used.

The detailed protocol extensions and procedures for such in-band signaling applications are described in other documents.

5.2. Multicast VPN

A L3VPN service that supports multicast is known as a Multicast VPN, or MVPN for short. There have been different proposed messages, procedures and mechanisms to support MVPN. These methods differ in protocols used in the service provider's network, for example, the mGRE-based MVPN, BGP extensions to transport customer's PIM signaling and P2MP RSVP-TE extensions to transport multicast data streams, and mLDP-based MVPN.

The receiver-driven multicast extensions to RSVP-TE can be used to support multicast VPN. Such an approach will greatly reduce the number of trees and multicast states in the core compared with the P2MP RSVP-TE approach.

The detailed procedures and mechanisms are described in [I-D.hlj-l3vpn-mvpn-mrsvp-te].

6. Fast Re-Route Considerations

The Fast Re-Route mechanisms and procedures specified in [RFC 4090] will not be applicable to the receiver-driven extension to RSVP-TE described in this document, since their Path/Resv messages are sent in different directions.

Extensions to mRSVP-TE to support Fast Re-Route are described in the document [I-D.zlj-mpls-mpls-mrsvp-te-frr].

7. Backward Compatibility

A receiver-driven P2MP LSP mechanism uses different C-Types than those in the sender-driven P2MP RSVP-TE. If LSRs do not recognize the receiver-driven C-Types, they will not support the receiver-driven extensions described in this document. LSRs that do not support receiver-driven P2MP-TE LSP, send Path Error [TBD] back to the Path Originator.

The complete discussion on the backward compatibility will be provided in the Next version of the document.

8. Acknowledgements

We would like to thank Lin Han and Katherine Zhao for their comments on early drafts of this work. In particular we would like to thank Lou Berger and Eric Osborne for their very helpful questions, comments and suggestions on our presentation of this work in Paris.

9. IANA Considerations

This section is TBD.

10. Security Considerations

How a receiver is authenticated is outside the scope of this document. But we will briefly summarize the requirements which are detailed in the requirements draft.

It is a requirement that any mRSVP-TE solution developed to meet some or all of the requirements expressed in this document MUST include mechanisms to enable the secure establishment and management of mRSVP-TE MPLS-TE LSPs. This includes, but is not limited to:

- o A receiver MUST be authenticated before it is allowed to establish mRSVP-TE LSP with its source, in addition to hop-by-hop security issues identified by in RFC 3209 and RFC 4206.
- o mechanisms to ensure that the ingress LSR of a P2MP LSP is identified;
- o mechanisms to ensure that communicating signaling entities can verify each other's identities;
- o mechanisms to ensure that control plane messages are protected against spoofing and tampering;
- o mechanisms to ensure that unauthorized leaves or branches are not added to the mRSVP-TE LSP; and
- o mechanisms to protect signaling messages from snooping.
- o Note that mRSVP-TE signaling mechanisms built on P2P RSVP-TE signaling are likely to inherit all the security techniques and problems associated with RSVP-TE. These problems may be exacerbated in mRSVP-TE situations where security relationships may need to be maintained between an ingress LSR and multiple egress LSRs. Such issues are similar to security issues for IP

multicast.

- o It is a requirement that documents offering solutions for P2MP LSPs MUST have detailed security sections.

11. References

11.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC2702] Awduche, D., Malcolm, J., Agogbua, J., O'Dell, M., and J. McManus, "Requirements for Traffic Engineering Over MPLS", RFC 2702, September 1999.
- [RFC3031] Rosen, E., Viswanathan, A., and R. Callon, "Multiprotocol Label Switching Architecture", RFC 3031, January 2001.
- [RFC3209] Awduche, D., Berger, L., Gan, D., Li, T., Srinivasan, V., and G. Swallow, "RSVP-TE: Extensions to RSVP for LSP Tunnels", RFC 3209, December 2001.
- [RFC4461] Yasukawa, S., "Signaling Requirements for Point-to-Multipoint Traffic-Engineered MPLS Label Switched Paths (LSPs)", RFC 4461, April 2006.
- [RFC4875] Aggarwal, R., Papadimitriou, D., and S. Yasukawa, "Extensions to Resource Reservation Protocol - Traffic Engineering (RSVP-TE) for Point-to-Multipoint TE Label Switched Paths (LSPs)", RFC 4875, May 2007.
- [RFC4420] Farrel, A., Papadimitriou, D., Vasseur, J., and A. Ayyangar, "Encoding of Attributes for Multiprotocol Label Switching (MPLS) Label Switched Path (LSP) Establishment Using Resource Reservation Protocol-Traffic Engineering (RSVP-TE)", RFC 4420, February 2006.
- [RFC4206] Kompella, K. and Y. Rekhter, "Label Switched Paths (LSP) Hierarchy with Generalized Multi-Protocol Label Switching (GMPLS) Traffic Engineering (TE)", RFC 4206, October 2005.
- [RFC2205] Braden, B., Zhang, L., Berson, S., Herzog, S., and S. Jamin, "Resource ReSerVation Protocol (RSVP) -- Version 1 Functional Specification", RFC 2205, September 1997.
- [RFC4090] Pan, P., Swallow, G., and A. Atlas, "Fast Reroute

Extensions to RSVP-TE for LSP Tunnels", RFC 4090, May 2005.

- [RFC3471] Berger, L., "Generalized Multi-Protocol Label Switching (GMPLS) Signaling Functional Description", RFC 3471, January 2003.

11.2. Informative References

- [I-D.zlj-mpls-mrsvp-te-frr]
Zhao, K., Li, R., and C. Jacquenet, "Fast Reroute Extensions to Receiver-Driven RSVP-TE for Multicast Tunnels", draft-zlj-mpls-mrsvp-te-frr-00 (work in progress), July 2012.
- [I-D.hlj-l3vpn-mvpn-mrsvp-te]
Han, L., Li, R., and C. Jacquenet, "Multicast VPN Support by Receiver-Driven Multicast Extensions to RSVP-TE", draft-hlj-l3vpn-mvpn-mrsvp-te-00 (work in progress), July 2012.
- [RFC3468] Andersson, L. and G. Swallow, "The Multiprotocol Label Switching (MPLS) Working Group decision on MPLS signaling protocols", RFC 3468, February 2003.
- [RFC3473] Berger, L., "Generalized Multi-Protocol Label Switching (GMPLS) Signaling Resource ReserVation Protocol-Traffic Engineering (RSVP-TE) Extensions", RFC 3473, January 2003.
- [RFC3564] Le Faucheur, F. and W. Lai, "Requirements for Support of Differentiated Services-aware MPLS Traffic Engineering", RFC 3564, July 2003.
- [RFC5467] Berger, L., Takacs, A., Caviglia, D., Fedyk, D., and J. Meuric, "GMPLS Asymmetric Bandwidth Bidirectional Label Switched Paths (LSPs)", RFC 5467, March 2009.

Authors' Addresses

Renwei Li
Huawei Technologies
2330 Central Expressway
Santa Clara, CA 95050
USA

Email: renwei.li@huawei.com

Quintin Zhao
Huawei Technologies
Boston, MA
USA

Email: quintin.zhao@huawei.com

Christian Jacquenet
France Telecom Orange
4 rue du Clos Courtel
35512 Cesson Sevigne,,
France

Email: christian.jacquenet@orange-ftgroup.com

MPLS Working Group
Internet-Draft
Intended status: Standards Track
Expires: October 7, 2012

P. Dutta
M. Aissaoui
Alcatel-Lucent
April 05, 2012

Multiple LDP Instances
draft-pdutta-mpls-multi-ldp-instance-00

Abstract

This document defines an extension to Label Distribution Protocol (LDP) [RFC5036] for implementation of multiple LDP instances in a network node, where all such instances share the common data plane. Multiple LDP instances provide a method for operators for fate separation of various LDP FEC Types as well as for network segmentation. The methods defined in this extension are backward compatible with procedures defined in [RFC5036]

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 October 7, 2012.

Copyright Notice

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

This document is subject to BCP 78 and the IETF Trust's Legal

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

Table of Contents

1. Introduction	3
2. Multiple LDP Instances	4
2.1. Procedures for multi-instance peering	4
2.1.1. Case 1	5
2.1.2. Case 2	6
2.1.3. Case 3	7
2.1.4. Case 4	8
3. Detection of multi-instance peering	8
4. LDP Address Distribution with multi-instance peering	9
5. LDP State Sharing between instances	9
6. Applicability	9
7. IANA Considerations	9
8. Security Considerations	10
9. Acknowledgements	10
10. References	10
10.1. Normative References	10
10.2. Informative References	10
Appendix A. An Appendix	11
Authors' Addresses	11

1. Introduction

The Multi-Protocol Label Switching (MPLS) architecture is described in [RFC3031]. Label Distribution Protocol (LDP) is a signaling protocol for setup and maintenance of MPLS LSPs (Label Switched Paths) and the protocol specification is defined in [RFC5036].

Two Label Switched Routers (LSR) that use LDP to exchange label/FEC mapping information are known as "LDP Peers" with respect to that information, and we speak of there being an "LDP Session" between them. A single LDP session allows each peer to learn the other's label mappings. Each LSR is identified by an LDP identifier. An LDP Identifier is a six octet quantity used to identify an LSR label space. The 4 octets identify the LSR and is a globally unique value, such as a 32-bit router Id assigned to the LSR. The last two octets identify a specific label space within the LSR. The last two octets of LDP Identifiers for platform-wide label spaces are always both zero. This document uses the following representation for LDP Identifiers:

<LSR Id> : <label space id>

e.g, lsrl71:0, lsrl9:2 etc

As per [RFC5036] an LSR that manages and advertises multiple label spaces uses a different LDP Identifier for each such label space. This means for a single label space there can be only one router-id that is can be assigned to the node that exclusively owns that label space. For example, it is not possible to have two LSRs like lsrl00:0 and lsr200:0 to be created in the a single node.

A LDP peering session between two LSRs may exchange labels for setting up LSPs that may belong to different FEC types. Operators may need the flexibility for fate separation of different FEC types in LDP protocol signaling when all such fec types share the same common label space. This is not possible with the current paradigm of single peering session between two LSRs and it requires one session per fate separated group of FEC types to exchange labels. Thus multiple LDP sessions are required between two peering nodes. One example could be fate separation between IP transport network and the overlay network of Pseudowires (PW). Procedures for PW set-up and maintenance using LDP are defined in [RFC4447]. It may be also desirable for fate separation IPv4 and IPv6 LSP set-up and maintenance in LDP in case of which two separate LDP sessions need to be formed between two peering nodes.

Although [RFC5036] does not specify that the 4 byte router-id of the LDP identifier be routable IP addresses, for various operational

simplicity implementations may map the 32 bit router-id to a IPv4 address configured in the node which is routable. In that way uniqueness of the 4 byte router-id can be achieved over a single routing domain. Interior Gateway Protocols (IGPs) like OSPF provide the option of creation of multiple instances for segmentation of a network into multiple routing domains. When LDP is deployed in such networks it is required to segment LDP network to align with multiple routing domains. When a node is connected to multiple such domains, LDP peering sessions over all such domains cannot use a common IPv4 router-id which is local to that node, since the IPv4 mapped router-id may not be routable across all such domains for security purposes. There are applications such as BGP Autodiscovery of L2VPNs or Dynamic MS-PW set-up that may auto-instantiate Targeted LDP sessions where BGP IPv4 next-hop addresses for respective NLRI's are mapped to peer LDP identifiers. Such next-hop addresses may not be routable between two routing domains. Thus there is need to host multiple LSRs by a network node that shares the same label space but each with unique router-ids.

This document describes a method to implement multiple instances of LDP in a network node that shares same label space. The method is generic and is backward compatible with nodes that supports procedures defined in [RFC5036] but does not support the procedures defined in this document. The procedures defined in this document would be referred as "Multi-Instance LDP".

2. Multiple LDP Instances

The solution defines the concept of implementing multiple LDP instances on a single network node that shares the single label space and thus shares the common data plane. Each such LDP instance is identified by a unique 4 byte router-id but same label space. Since the multi-instance procedures use same LDP Identifier as defined in [RFC5036], it makes the node running multiple instances to be backward compatible with the node that support the multi-instance LDP procedures.

2.1. Procedures for multi-instance peering

When multiple LDP instances are set-up between two peering nodes for fate separation reasons then there can be various ways Hello adjacencies can be formed over the interfaces between the nodes. Further multi-instance peering for fate separation results in multiple parallel sessions between two peering nodes.

While running parallel multi-instance LDP sessions between two peering nodes,

1. Each peering session MUST use separate transport address.
2. The FEC label mappings exchanged over each peering session MUST be a disjoint set from one another.

The above rules does not apply between multi-instance LDP sessions with different peering LDP nodes.

This document describes the following cases and defines the rules and procedures with each case.

2.1.1. Case 1

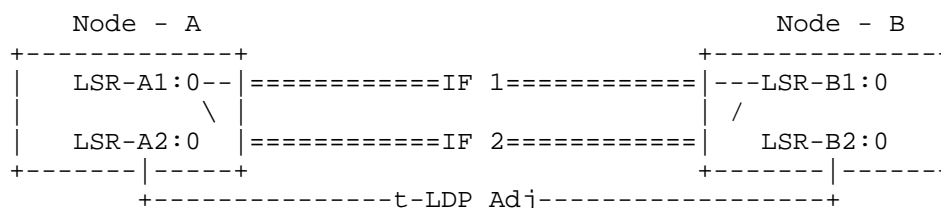


Figure 1.

In this case the operator wants to separate the fate of FECs exchanged between the nodes into two separate groups - Group 1 and Group 2. For example Group 1 can contain all transport specific FEC types such as IPV4 FEC Element Type and LDP Multi-point (MP) FEC types etc. LDP Multi-point FEC types are described in [RFC6388]. Group 2 contains various Pseudowire (PW) FEC types. PW setup and maintenance using LDP is described in [RFC4447].

Two separate LSR-IDs are provisioned in each node - one LSR is dedicated for FEC Group 1 and another for FEC Group 2.

There are two parallel interfaces between Node-A and Node-B as IF1 and IF2 respectively.

The traffic for LSPs set-up for FEC Group 1 may use both IF1 and IF2. Thus both IF1 and IF2 would exchange Hello Packets using LSR-A1:0 and LSR-A2:0 for setting up Hello adjacency for the LDP instance assigned for FEC Group 1. The Hello messages exchanged over IF1 and IF2 MUST carry the LDP Adjacency Capabilities for each FEC Types in FEC Group 1. LDP Adjacency Capabilities are defined in [LDP-ADJ-CAP]. This would result in formation of a LDP session between Node A and Node B for the instance identified by LSR-A1:0 and LSR-B1:0 respectively. The LDP session SHOULD be set-up with Capabilities of FEC Group 1.

LDP session specific capability negotiation is described in [RFC5561]

A Targeted LDP (t-LDP) hello adjacency would be formed between node A and node B using LSR-A2:0 and LSR-B2:0 respectively. The t-Ldp Hello Messages exchanged between the nodes MUST carry the LDP Adjacency Capabilities for each FEC Types in FEC Group 2. This would result in a LDP session between Node A and Node B for the instance identified by LSR-A2:0 and LSR-B2:0 respectively. The LDP session SHOULD be set-up with capabilities of FEC Group 2.

2.1.2. Case 2

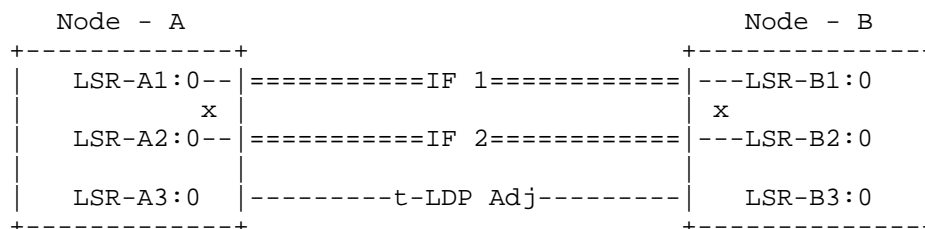


Figure 2.

This is a variant of case 1 where the operator may choose to further separate the fate of IPV4 FEC Element Type and MP FEC Element types into "Unicast" and "Multicast" Groups. Thus there are three FEC Groups here and fate separation is required for all three FEC Groups.

FEC Group 1 : IPv4 FEC Element Type.

FEC Group 2: MP FEC Element Types.

FEC Group 3: PW FEC Element Types.

LDP Instance 1: The LDP instance with peering LSR-A1:0 and LSR-B1:0 are assigned for FEC Group 1.

LDP Instance 2: The LDP Instance with peering LSR-A2:0 and LSR-B2:0 are assigned for FEC Group 2.

LDP Instance 3: The LDP instance with peering LSR-A3:0 and LSR-B3:0 are assigned for FEC Group 3.

In this case, both IF1 and IF2 are associated with LDP instances 1 and 2. Each of IF1 and IF2 would originate two separate Hello Messages using the same source IP address, one Hello Message for each

instance . This would result in two hello adjacencies per interface - one for Instance 1 and Instance 2. Each Hello Adjacencies SHOULD advertise capabilities using rules described in case 1.

Such case may also arise when operator wants to do fate separation of IPV4 and IPV6 LDP based LSPs but IF1 and IF2 are single stack interfaces only - that is either IPV4 or IPV6. Thus an operator may provision single stack interfaces IF1 and IF2 and yet can provision fate separation of IPV4 and IPV6 LSPs.

The t-Ldp Hello Adjacency would be formed for LDP Instance 3 using the PW Capabilities.

2.1.3. Case 3

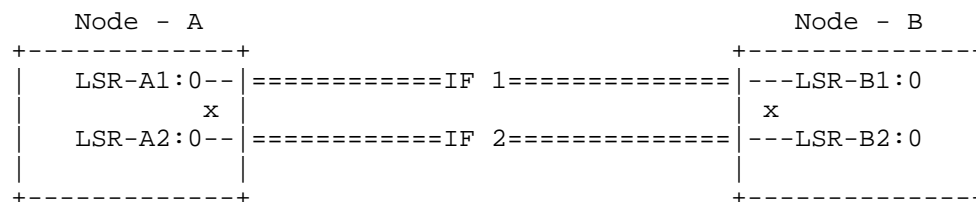


Figure 3.

This case is a variant of case 2 where, both interfaces IF1 and IF2 are dual-stack (IPV4 and IPV6) interfaces and operator wants fate separation of IPV4 and IPV6 LSPs. Without loss of generality, hereby IPV4 or IPV6 FECs may include all FEC types that are associated with IPV4 or IPV6. For example, [I-D.ietf-mppls-mldp-in-band-signaling] defines several in-band MP FEC Types that may be classified into IPV6.

LDP Instance 1: The LDP instance with peering LSR-A1:0 and LSR-B1:0 are assigned for IPV4 FEC Types.

LDP Instance 2: The LDP Instance with peering LSR-A2:0 and LSR-B2:0 are assigned for IPV6 FEC Types.

Both the interfaces IF1 and IF2 are associated with each of the LDP instances 1 and 2 respectively. Here the operator may choose to use IPV4 addresses on the interfaces for sending Hello Messages for Instance 1 and IPV6 addresses on the interfaces for sending Hello Messages for Instance 2.

2.1.4. Case 4

This case is variant of case 3 where interface IF1 is dedicated for IPV4 LSP Types and IF2 is dedicated for IPV6 LSP Types. This provides fate-separation of both control plane and data plane for LSP types.

3. Detection of multi-instance peering

While running parallel multi-instance LDP sessions between two peering nodes, it is important to detect that such sessions with the same peer node. If a node receives the same FEC label mapping from parallel multi-lsr peering sessions it may result in a loop for some applications. An example of such application can be LDP based Virtual Private LAN Service (VPLS) described [RFC4762]. So it is important to detect and prevent such loops.

This document defines a new LDP Node-ID TLV that uniquely identifies the node that hosts multiple LDP instances. The LDP Node ID TLV is OPTIONAL and is carried in LDP Hello Messages sent out by the node in its Optional Parameters. The encoding of the LDP Node ID TLV is as follows:

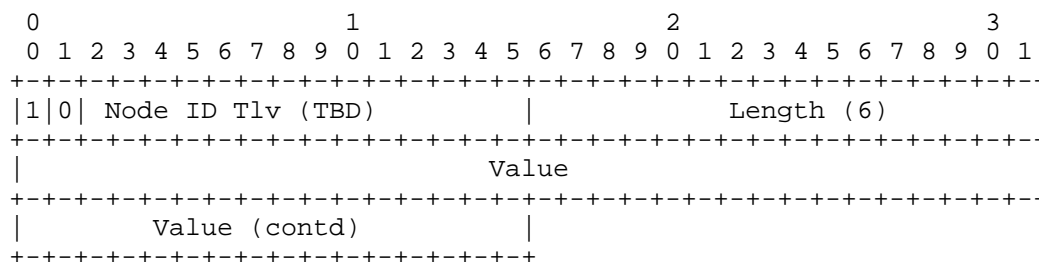


Figure 4

The Value field is a 48 bit identifier and MUST be unique identifier across the network.

1. All the multi-instance LDP LSRs MUST advertise the same LDP Node-ID TLV in all Hello Messages originated by that node. One example of the value can be a IEEE Vendor specific MAC Address that can uniquely identify a node in the network.

2. When a LSR receives a FEC label mapping from a peering session but same FEC mapping has been already receiver over another peering session associated with same Node-ID then the receiving LSR MUST send a Label Release to the peering session with statuc code

LOOP_DETECTED.

4. LDP Address Distribution with multi-instance peering

An LSR maintains learned labels in a Label Information Base (LIB). When operating in Downstream Unsolicited mode, the LIB entry for an address prefix associates a collection of (LDP Identifier, label) pairs with the prefix, one such pair for each peer advertising a label for the prefix. When the next hop for a prefix changes, the LSR retrieves the label advertised by the new next hop from the LIB for use in forwarding. To retrieve the label, the LSR should be able to map the next hop address for the prefix to an LDP Identifier. Similarly, when the LSR learns a label for a prefix from an LDP peer, it should be able to determine whether that peer is currently a next hop for the prefix to determine whether it needs to start using the newly learned label when forwarding packets that match the prefix. To make that decision, the LSR should be able to map an LDP Identifier to the peer's addresses to check whether any are a next hop for the prefix. To enable LSRs to map between a peer LDP Identifier and the peer's addresses, LSRs advertise their addresses using LDP Address and Withdraw Address messages as per procedures defined in [RFC5036]

However while running multi-instance LDP peering between two nodes, it is possible that all such sessions would distribute same set of local addresses in each node. An implementation MAY segregate the local address space in each node among the multiple ldp instances to avoid duplication of address distribution.

5. LDP State Sharing between instances

TBD.

6. Applicability

This solution described in this document is applicable for multi-instance LDP sessions for fate separation as well as for segmentation of LDP network domains. More details would be covered in next revisions of the document.

7. IANA Considerations

This document requests the following code points:

- LDP Node-ID TLV type.

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

8. Security Considerations

[I-D.ietf-mpls-mpls-and-gmpls-security-framework] describes the security framework for MPLS networks. whereas [RFC5036] describes the security considerations that apply to the base LDP specification. The same security framework and considerations apply to the capability mechanism described in this document.

9. Acknowledgements

The authors would like to thank Wim Henderickx for insightful comments and probing questions.

10. References

10.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC3031] Rosen, E., Viswanathan, A., and R. Callon, "Multiprotocol Label Switching Architecture", RFC 3031, January 2001.
- [RFC5036] Andersson, L., Minei, I., and B. Thomas, "LDP Specification", RFC 5036, October 2007.
- [RFC5561] Thomas, B., Raza, K., Aggarwal, S., Aggarwal, R., and JL. Le Roux, "LDP Capabilities", RFC 5561, July 2009.
- [RFC6388] Wijnands, IJ., Minei, I., Kompella, K., and B. Thomas, "Label Distribution Protocol Extensions for Point-to-Multipoint and Multipoint-to-Multipoint Label Switched Paths", RFC 6388, November 2011.

10.2. Informative References

- [I-D.ietf-mpls-mldp-in-band-signaling] Wijnands, I., Eckert, T., Leymann, N., and M. Napierala, "Multipoint LDP in-band signaling for Point-to-Multipoint and Multipoint-to-Multipoint Label Switched Paths",

draft-ietf-mpls-mldp-in-band-signaling-05 (work in progress), December 2011.

[I-D.ietf-mpls-mpls-and-gmpls-security-framework]

Fang, L. and M. Behringer, "Security Framework for MPLS and GMPLS Networks",
draft-ietf-mpls-mpls-and-gmpls-security-framework-09 (work in progress), March 2010.

[RFC4447] Martini, L., Rosen, E., El-Aawar, N., Smith, T., and G. Heron, "Pseudowire Setup and Maintenance Using the Label Distribution Protocol (LDP)", RFC 4447, April 2006.

[RFC4762] Lasserre, M. and V. Kompella, "Virtual Private LAN Service (VPLS) Using Label Distribution Protocol (LDP) Signaling", RFC 4762, January 2007.

Appendix A. An Appendix

Authors' Addresses

Pranjal Kumar Dutta
Alcatel-Lucent
701 E Middlefield Road
Mountain View, California 94043
USA

Phone:

Fax:

Email: pranjal.dutta@alcatel-lucent.com

Mustapha Aissaoui
Alcatel-Lucent
600 May Road
Kanata, ON
Canada

Phone:

Fax:

Email: mustapha.aissaoui@alcatel-lucent.com

URI:

INTERNET-DRAFT
Intended Status: Standards Track
Expires: Expires January 10, 2013

B. Tao, Ed.
Huawei Technologies
Others
July 9, 2012

MPLS PIM Inter-working
draft-cao-mpls-pim-interworking-00

Abstract

This document describes a framework for the inter-working between Protocol Independent Multicast [PIM] and a leaf-driven P2MP tunnel signaling protocol such as [mRSVP-TE] or [mLDP] so that multiple PIM sites around an MPLS network can form a single PIM domain without compromising PIM's features, scalability, and performance.

In this document, PIM modes PIM-SM, PIM-SSM, and PIM-BIDIR are considered.

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

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

Table of Contents

1	Introduction	4
1.1	Background	4
1.2	Purpose of This Work	5
1.3	Terminology	6
2.	An Overview: PIM MPLS Inter-working	6
2.1	PIM-SM, PIM-SSM and PIM-Bidir States	6
2.2	PIM-MPLS Border Router(mPMBR)	7
2.3	A Reference Model for PIM-MPLS Inter-working(PMIW)	9
2.3.1	QPI, MPLS Tunnel and M-Flow Spec Binding	10
2.3.2	PIM Support for QPI and M-Flow Spec Binding	10
2.3.3	M-Flow Spec Binding Policies	11
2.3.4	IP Multicast Packet Forwarding at an mPMBR	11
2.4	Impacted PIM messages and procedures	11
2.4.1	PIM Hello and Adjacency Over MPLS Backbone	12
2.4.2	PIM Assert and Message	12
2.4.3	PIM hop-by-hop Bootstrapping and Message	12
2.4.4	PIM Unicast Messages and C-RP Advertisement	12
2.4.5	PIM RP Register and RegisterStop	13
2.4.6	PIM Join/Prune States and In-Band Signaling in MPLS	13
2.4.6.4	Aggregating Two Tunnels to Use A Single Tunnel	14
3	Algorithms and Procedures	14
3.1	Dynamic P2MP Tunnel Creation and Bind An M-Flow Spec To It	14
3.1.1.	In-Band Tunnel Signaling at Leaf LER	16
3.1.2.	In-Band Tunnel Signaling at Transit LSR	17
3.1.3.	In-Band Tunnel Signaling at Root LER	18
3.1.4.	Considerations for Load Balance and Traffic Engineering(TE)	18
3.2.	Operational Procedures for PIM RPT to SPT switch	19
4.	OAM & P	19
5	Security Considerations	20
6	IANA Considerations	20

7	References	20
7.1	Normative References	20
7.2	Informative References	20
	Authors' Addresses	21
	Acknowledgement	21

1 Introduction

1.1 Background

IP multicast data sources and their receivers can be located at multiple separated sites which are around an MPLS backbone. PIM, the most popular IP multicast protocol, runs in each of these sites. A requirement therefore is to use the MPLS backbone tunnels to "connect" these multiple PIM sites as if they were in a single multicast domain under PIM.

Currently there are a few standards to achieve this, most of them for multicast VPN(mVPN) cases:

- a) [RFC6513] and [RFC6514] use a third protocol, the extended BGP, to discover multicast routes and other states from each PIM site, propagate to other sites, and establish P2MP tunnels in the MPLS backbone to support VPN multicast within MPLS backbone.
- b) [I-D.lin-mrsvp-te-mvpn] uses an mRSVP-TE [mRSVP-TE] tunnel within MPLS to transparently multicast both PIM's control and data traffic to other VPN sites, and if necessary, establishes a separated P2MP tunnel for each individual multicast flow. This is an out-of-band method to signal VPN PIM states over MPLS backbone. In this way, separated PIM sites of a VPN form a single PIM domain in the VPN, and PIM adjacencies each pair of MPLS edge routers are maintained across the MPLS backbone.
- c) [I-D.Wijnands-mldp-inband] uses in-band signaling to build mLDP P2MP tunnels to support (S, G) and (*, G) multicast states. Currently, the draft only specifies the encoding and decoding for the above two types of multicast states. It relies on a third protocol to signal PIM ASM related states.

These methods, however, either support a limited portion of PIM features, or, with a concentration in mVPNs, have remaining issues for both network operators and equipment vendors.

[RFC6513] extends BGP to support PIM features cross over an MPLS backbone for multicast VPN cases, however, the support is made critically dependent on the extensions to the third protocol. Some of PIM features such as BSR bootstrapping are not supported yet. To fully support PIM and its future extensions, more extensions to BGP must be made as well, besides those for PIM and the MPLS protocols. This dependency causes additional requirements and complexity for

both network operators and equipment vendors.

The extra protocol involvement also introduces more interactions among protocols and thus causes additional overheads to BGP. In addition, [RFC6513] uses a BGP discovery phase for a PIM state before a tunnel can be signaled in the MPLS backbone. This adds a delay to the dynamic tunnel signaling.

The out-of-bound method limits itself as a solution for particular cases because: a) It applies to a particular MPLS signaling protocol [mRSVP-TE]; b) Fully meshed PIM adjacencies over MPLS are costly and can have scalability concern (See [RFC6517]); c) The default tunnel may not be optimally routed within MPLS and its usage prevents flexible load balancing and traffic engineering at tunnel level; only limited aggregation can be done on (S, G) data tunnels; d) PIM RPT to SPT switch semantic is changed and new procedures and messages are used to discover a (S, G) and propagate it to other sites.

In the third solution, the supported PIM features are limited. Besides, it applies only to [mLDP] as the MPLS signaling protocol.

See [RFC6517] for more information.

1.2 Purpose of This Work

In this document, we introduce a PIM-MPLS inter-working framework which provides the following to resolve the current issues:

- a) Complete the full support of PIM features with only PIM and a point-to-multipoint(P2MP) tunneling protocol involved;
- b) Neutrally support various tunnel signaling protocols, including [mRSVP-TE] and [mLDP]; PIM states are "in-band" signaled by the tunnel signaling protocol to another PIM site while the signaling protocol itself uses the data to set up the tunnel with optimal routing and traffic engineering;
- c) Minimize the changes to the two(2) involved protocols and the introduction of new procedures;
- d) Provide flexible tunnel aggregation and load balancing for scalability and shared resource usage in backbone
- e) Minimize overheads in the backbone and on the PIM-MPLS border routers without introducing performance and scalability bottlenecks. There is no PIM adjacencies cross over the backbone network

It is important to point out that some existing solutions can be made

to work with this framework to have complete PIM support.

[mRSVP-TE] and [mLDP] are protocols to signal point-to-multipoint(P2MP) and multipoint to multipoint(MP2MP) tunnels in an MPLS network, starting from the leaves of these tunnels. The leaf-driven signaling is in the same direction as PIM builds its multicast forwarding information base, i.e., from the multicast data listeners to the senders. This framework will take advantage of this characteristics to set up the forwarding states in an MPLS backbone with less messages and delays.

1.3 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. An Overview: PIM MPLS Inter-working

2.1 PIM-SM, PIM-SSM and PIM-BiDir States

[PIM] defines four(4) types of Multicast Routing Information Base(MRIB) entries:

1. (*, *, RP)
2. (*, G)
3. (S, G)
4. (S, G, rpt)

For each MRIB entry, the framework identifies the following PIM forwarding states for our purpose in this document:

Downstream Per-interface Join/Prune state:

One of {"NoInfo" (NI), "Join" (J)}

Upstream non-interface specific Join/Prune state:

One of {"NotJoined", "Joined"}

For each (*, G), there is a sub-set of states called (S, G, RPT), one for each (S, G) pair. In this draft, we are only concerned with their following states:

Downstream Per-interface Join/Prune state:

One of {"NoInfo", "Pruned"}

Upstream non-interface specific Join/Prune State:

One of {"RPTNotJoined(G)", "NotPruned(S,G,rpt)",
"Pruned(S,G,rpt)"}

For definitions of these states, readers are referred to the [PIM] document.

This framework makes these PIM states as part of signaling data for a leaf-driven MPLS protocol to signal its P2MP tunnels to achieve optimal multicast routing within the MPLS network and in highly scalable fashion. On the other hand, the remote PIM site at upstream obtains these states from the MPLS signaling data to set up proper PIM forwarding states for the upstream site. These PIM states are therefore "In-Band" signaled to the remote PIM site by MPLS, while MPLS itself sets up optimized multicast LSPs for the traffic to go through the MPLS backbone.

We hereafter call them M-Flow Specs, and for in-band signaling purpose, assign a value to each of the types as the following:

M-Flow Spec Type-1(value 1) for (*, *, RP);
M-Flow Spec Type-2(value 2) for (*, G);
M-Flow Spec Type-3(value 3) for (S, G);
M-Flow Spec Type-4(value 4) for (S, G, RPT);

2.2 PIM-MPLS Border Router(mPMBR)

An mPMBR is a border router where PIM meets the MPLS backbone and inter-works with an MPLS signaling protocol to set up the tunnels, and where IP multicast packets exit an upstream PIM site to enter the MPLS backbone or exit the MPLS backbone to enter a downstream PIM site. An mPMBR can run a multicast member discovery protocol such as

IGMP or MLD, besides PIM. Therefore the mPMBR can have local listeners.

An mPMBR is called local to a PIM site if it is where the PIM site connects to the MPLS backbone.

For convenience in the following discussions, we define the following macro to determine where a P2MP tunnel ingress, or root, will be, for each M-Flow spec F defined previously:

```
mPMBR_lkup(F) = {  
    RP's local mPMBR address, if F is a (*, *, RP) spec; or  
    RP(G)'s local mPMBR address if F is a (*, G) spec; or  
    S's local mPMBR address if F is a (S, G) spec; or  
    RP(G)'s local mPMBR if F is a (S, G, RPT)  
}
```

An mPMBR needs to specify its router ID and advertise it to unicast routing protocol(s) to make the address reachable from all other mPMBRs. It also specifies its PIM interfaces that face a PIM site, as well as one or more MPLS interfaces over which P2MP tunnels can be signaled to support the inter-working functions as defined in this work. In this framework, P2MP tunnels and their sub-LSPs are dynamically created and removed when M-Flow specs are created or removed on mPMBRs.

When a tunnel is created for an M-Flow spec, a logic interface is also created at the tunnel's ingress and egress endpoints, respectively. This logic interface will act as if it were a PIM interface but it does not actually run PIM on it. At an P2MP egress mPMBR, PIM builds non-interface upstream states on it using the M-Flow spec created by PIM; at the ingress, or P2MP root mPMBR, PIM builds per-interface downstream states using the M-Flow spec data signaled by the tunnel signaling protocol.

An M-Flow spec is said to be "bound" to such a logic interface once the interface is created as above to pass the traffic which will be forwarded per the M-Flow spec by the IP multicast forwarding plane. At a P2MP tunnel's ingress mPMBR, IP multicast packets of the bound M-Flow spec enters this logic interface, which "leads" the packets into the MPLS tunnel. At an egress mPMBR, the packets exit the tunnel and were treated as if they were received from the logic interface as an upstream interface. At this point the packets will be forwarded using the native IP multicast forwarding rules.

We call each of these logic interfaces a Quasi-PIM interface(QPI).

2.3 A Reference Model for PIM-MPLS Inter-working(PMIW)

Figure 1 illustrates the PIM-MPLS inter-working model on an mPMBR. In this model, a QPI is created for an MPLS tunnel at the ingress and egress LSRs, and this QPI is "advertised" to the mPMBR's PIM, so that PIM uses it as if it were a PIM interface except that PIM protocol does not actually run on it, and therefore PIM does not send or receive any PIM protocol control messages over it (i.e. there is no PIM adjacency on a QPI), but can still send and receive IP multicast data packets.

The PIM on each mPMBR uses native PIM procedures to work with its PIM site router(s) and build proper PIM control and multicast packet forwarding states over the PIM interfaces as well QPIs.

In order for the mPMBR PIM to build proper control and forwarding states for a QPI, PIM procedures must be modified to

- i) Extend any validation checks to include the QPI to accept IP multicast packets from the backbone;
- ii) PIM RFP_Interface() macro can return a QPI if the RPF next-hop goes over an IP interface that has MPLS enabled to support the inter-working.

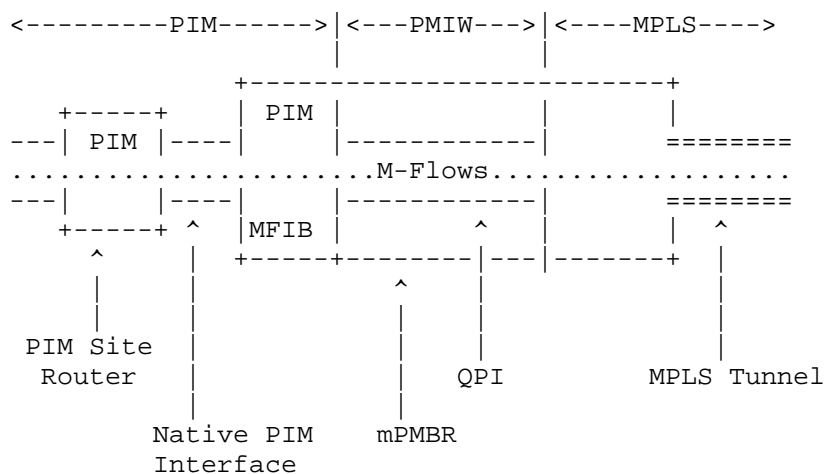


Figure 1: mPMBR PIM-MPLS Inter-working(PMIW) Reference Model

In the Figure 1, an IP multicast packet in mPMBR's PIM segment is forwarded over the PIM interface(s) as well as QPIs using PIM's native forwarding rules; when an IP multicast packet enters a QPI, the packet will be encapsulated with MPLS and enters the corresponding tunnel. When a packet exits the tunnel at the remote mPMBR, the QPI on this receiving mPMBR becomes the incoming interface for the packet, and the packet enters PIM's segment where it will be forwarded under PIM's native rules.

2.3.1 QPI, MPLS Tunnel and M-Flow Spec Binding

When a P2MP tunnel is created in the MPLS backbone for an M-Flow spec, a QPI is also created as an IP multicast logical interface at each of the egress and ingress mPMBRs, and the M-Flow spec is bound to the QPIs at each of the mPMBRs respectively. At the ingress mPMBR, the QPI is where an IP multicast packet enters the P2MP tunnel with the MPLS header, and is subsequently forwarded along the tunnel. At each egress mPMBR, the QPI associated with the tunnel is where the MPLS packet is de-capsulated, and the resulted IP multicast packet continues to be forwarded using PIM's native forwarding rules.

The QPI operational status is the same as the P2MP tunnel operational state.

At an ingress mPMBR, an M-Flow spec F is said to be "bound" to a QPI (therefore the tunnel as well) when the ingress mPMBR sets the IP multicast forwarding rules of F to use the QPI as a downstream interface in order to carry the traffic of F to other PIM sites, via the backbone network.

At an egress mPMBR, an M-Flow spec F is said to be "bound" to a QPI (therefore the tunnel as well) when the egress mPMBR sets the IP multicast forwarding rules of F to use the QPI as an upstream interface in order to receive the traffic of F from another PIM site, via the backbone network.

2.3.2 PIM Support for QPI and M-Flow Spec Binding

In this framework, QPIs are made to PIM as if they were a PIM interface, except that PIM control messages do not go over the QPIs. The implementation needs to establish proper PIM per-interface downstream states and upstream states for the QPI and the data signaled by MPLS, which is originated from other PIM interface states and the MPLS signaled data from the remote PIM sites.

The actual implementation of the binding on each mPMBR is beyond the scope of this work.

The detailed PIM protocol support for the QPI will be completed in a later version.

2.3.3 M-Flow Spec Binding Policies

This framework introduces two sets of policies to restrict the binding of an M-Flow to a Tunnel.

1. Default Policy: AGG_POLICY_0 (value: 0)
 - a. One tunnel per (S, G) M-Flow spec; and
 - b. One tunnel per RP for (*, *, RP) M-Flow spec; and
 - c. One tunnel for all M-Flow specs of (*, G) such that RP(G) gives the same RP

The default policy is used by a LSR when no other policy is explicitly specified. It is the finest grained, and MUST be provided by an implementation.

2. AGG_POLICY_1 (value: 1):
 - a. A separate P-Tunnel is used to aggregate all (S, G) M-Flow specs such that mPMBR_lkup((S,G)) gives the same mPMBR; and
 - b. A separate P-Tunnel is used to aggregate all (*, *, RP) M-Flow specs such that mPMBR_lkup((* , *,RP)) gives the same mPMBR; and
 - c. A separate P-Tunnel is used to aggregate all (*, G) M-Flow specs such that mPMBR_lkup((* , G)) gives the same mPMBR.

AGG_POLICY_1 is the most coarse grained and it, besides others, can be optionally provided by an implementation.

2.3.4 IP Multicast Packet Forwarding at an mPMBR

If an IP multicast packet arrives at an mPMBR from a PIM interface, it is forwarded using native IP multicast rules. If an oif is a QPI, the QPI makes the packet to be forwarded into the corresponding MPLS P2MP tunnel.

If an IP multicast packet arrives at an egress mPMBR from a P2MP tunnel, it is handled by the bound QPI, which acts as an iif for IP multicast flow. If there is no bound QPI, the packet is dropped.

2.4 Impacted PIM messages and procedures

2.4.1 PIM Hello and Adjacency Over MPLS Backbone

An mPMBR does not build any PIM adjacency with any other mPMBR over the MPLS backbone. Instead, relevant PIM states are mapped to and from the MPLS signaling data. The details will be covered in later sections when actual procedures are provided. The PIM downstream per-interface states on a QPI is directly mapped from the corresponding MPLS P2MP tunnel's in-band signaled M-Flow spec data. The PIM upstream states, on the other hand, will be in-band signed to other PIM sites by a P2MP tunneling protocol, and at the same time the signaling protocol sets up optimally routed P2MP tunnel within the MPLS for the multicast flows.

There are no impacts to other routers within MPLS backbone or in PIM sites.

2.4.2 PIM Assert and Message An implementation following this framework will not need to make any changes to for PIM asserts, as they will be only applicable inside a PIM site locally, and the framework does not let PIM go cross the backbone.

There are no impacts to any router in MPLS backbone or in PIM sites.

2.4.3 PIM hop-by-hop Bootstrapping and Message

A designated multicast channel within MPLS backbone, called Multicast Bootstrap Tunnel(MBT), is used to carry PIM's bootstrap messages among all mPMBRs. This tunnel can be either an MP2MP or a bi-directional P2MP tree. The root is designated by the MPLS network operator and leaves are all mPMBRs. An MPLS packet entered into the MBT from any mPMBR will reach all other mPMBRs.

At each mPMBR, PIM sends and receives bootstrap messages to each of the mPMBR's PIM interfaces as well as the MBT. Each mPMBRs must implement the functions to send and receive PIM bootstrap messages over the MBT.

There are no other impacts to an mPMBR PIM's native bootstrap procedures, and there are no impacts to other routers other than the mPMBRs.

2.4.4 PIM Unicast Messages and C-RP Advertisement

PIM unicast messages, including CRP advertisement, Register and RegisterStop, are sent and received in raw IP, with PIM protocol number.

Each mPMBR must be able to receive a raw IP PIM packet arrived at a non-PIM interface that is MPLS enabled to support PIM-MPLS inter-working.

There are no other impacts to PIM's native procedures for unicast messages on an mPMBR, and there are no impacts to other routers other than mPMBRs.

2.4.5 PIM RP Register and RegisterStop

Except for the changes common to PIM unicast messages as in the previous subsection, there are no other impacts for PIM RP register and registerStop.

For information purpose, a RP may choose to send a (S, G) join toward the source S after an (S, G) register packet is received by the RP, and the resulted tree may go over the MPLS network. In this case, the mechanism and procedures for (S, G) SPT support apply. The details will be in later subsections.

2.4.6 PIM Join/Prune States and In-Band Signaling in MPLS

An mPMBR, say `mpmbr`, can have four(4) types of M-Flow specs corresponding to PIM's upstream states. Except for the M-Flow spec Type-4, each of the rest, say, `F` at the `mpmbr` can bind to an MPLS P2MP tunnel in the MPLS network with `mpmbr` as one of its leaf(egress) LSRs, and the root set to `mPMBR_lookup(F)`. A signaling protocol which is to work under this framework will support the binding of the Type-1, Type-2, and Type-3 M-Flow specs with its tunnels. A Type-4 M-Flow spec is associated with a Type-2 M-Flow spec, as an (S, G, RPT) state is embedded into a (*, G) state. Therefore there is no separate binding for a Type-4 M-Flow spec.

Before an M-Flow spec `F` is bound to a tunnel, the to-be bound tunnel may have some undetermined information such as a tunnel identifier, but the tunnel signaling procedure uses `F` and other known tunnel identification data during the tunnel signaling. After the M-Flow spec is bound to a tunnel, tunnel's identification data will be completed.

The binding can happen at the leaf, at a transit LSR, or at the root, according to the binding procedure in "Algorithms and Procedures".

2.4.6.4 Aggregating Two Tunnels to Use A Single Tunnel

Two tunnels may be aggregated into a single one, if their M-Flow specs can be merged to form a super set of M-Flow specs, without violating each original tunnel's aggregation policy. The policy tests are performed using the algorithms and procedures as defined in Section 3.

The merged tunnel can be set up using Make-Before-Break(MBB). They must have the same root in order to be aggregated. The procedure of P2MP MBB is beyond the scope of this draft.

3 Algorithms and Procedures

3.1 Dynamic P2MP Tunnel Creation and Bind An M-Flow Spec To It

This section describes the procedure to bind an M-Flow spec to a tunnel.

First, each M-Flow spec F has an aggregation policy to restrict aggregating the M-Flow spec into a tunnel. This policy can be configured explicitly by an operator, or is the default policy if it is not configured.

An M-Flow spec F has the following attributes:

F.agg_policy:	An policy to restrict binding and aggregating F into a tunnel. This policy can be configured explicitly by an operator, or is the default policy if it is not configured.
F.spec_type:	One of the spec types.
F.bound_tnl:	The MPLS tunnel used by the IP multicasting data which enters and exits the tunnel per F's corresponding PIM forwarding rules.

For simplicity purpose, and without implying actual implementations, the framework uses the following macros and functions on each LER or LSR:

```
TS = {all tunnels on a node that supports PIM-MPLS
      inter-working};
mflow_specs(T) = {set of M-Flow specs that have bound to
                  tunnel T}
mflow_spec_type(T) = The type of the M-Flow specs that are
                     bound to tunnel T
```

```
agg_policy_compatible(F, T) = TRUE if and only if
    T.agg_policy derives F.agg_policy

agg_policy_compatible(T1, T2) = TRUE if and only if
    T1.agg_policy derives T2.agg_policy

bind_candidate(F) {
    foreach(T in TS) {
        if ((T' = F.bound_tnl) != NULL && T' signaling is
            completed)
        {
            return T';
        }
        if (F.spec_type == mflow_spec_type(T) &&
            agg_policy_compatible(F, T))
        {
            return T;
        }
    }
    return NULL;
}

mflow_spec_bind(F, LSR) {
    if ((T = bind_candidate(F)) != NULL)
    {
        mflow_specs_merge(T);
        F.bound_tnl = T;
    }
    else if (I_am_leaf(LSR))
    {
        T = initiate_signal_p2mp_tunnel(F);
        mflow_specs(T) = {F};
        F.bound_tnl = T;
    }
    else if (I_am_root(LSR))
    {
        T = continue_signal_p2mp_tunnel(F);
        mflow_specs(T) = {F};
        F.bound_tnl = T;
    }
    else //I am transit LSR
    {
        T = continue_signal_p2mp_tunnel(F);
        mflow_specs(T) = {F};
    }
}
```

init_signal_p2mp_tunnel(F) triggers a P2MP tunnel creation using the local LER as the leaf, and mPMBR_lkup(F) as the root.

continue_signal_p2mp_tunnel(F) continues the signaling procedure for the P2MP tunnel that was initiated for F.

The following defines M-Flow spec merge operation:

```
mflow_specs_merge(T, F)
{
    switch(F.spec_type)
    {
        case Type-1:
            /* mflow_specs(T) is a set of group ranges each with
               a list of (S, G, RPT) entries; F must be a group
               range with a list of its (S, G, RPT) entries */
            mflow_specs(T) = mflow_specs(T) union {F};
            break;
        case Type-2:
            mflow_specs(T).sg_rpt_joins =
                mflow_specs(T).sg_rpt_joins union F.sg_rpt_joins;
            mflow_specs(T).sg_rpt_prunes =
                mflow_specs(T).sg_rpt_prunes intersect
                F.sg_rpt_prunes
            /* sg_rpt_joins and sg_rpt_prunes are the lists of G's
               joined and pruned(S, G, RPT) entries */
            break;
        case Type-3:
            mflow_specs(T) = mflow_specs(T) union {F};
            /* mflow_specs(T) is a set of (S, G) entries */
            break;
        case Type-4:
            break; /* (S, G, RPT) entries go with wildcards */
    }
}
```

The actual procedures for initiate_signal_p2mp_tunnel(F) and continue_signal_p2mp_tunnel(F) are MPLS signaling protocol specific and will be out of scope for this document.

3.1.1. In-Band Tunnel Signaling at Leaf LER

An M-Flow spec F is instantiated when an mPMBR PIM creates a J/P upstream state. There are three cases under which F may be bound to a

P2MP tunnel at the leaf LER:

Case 1: F is an M-Flow spec already bound to a tunnel egress. Therefore the corresponding QPI of the tunnel is used for F.

Case 2: F is not bound to an existing tunnel yet but F's binding policy is compatible with an M-Flow spec which has been bound to a P2MP tunnel. Therefore, F is then bound to the same tunnel, and its QPI is used for F. The leaf mPMBR does not initiate a new tunnel.

Case 3: None of Case 1 and Case 2 is true. Then a tunnel with some unknown identifier is signaled using an extended MPLS protocol, and F as additional data for in-band signaling. Binding does not happen at this time.

After the unknown tunnel signaling is completed, an upstream node, either a transit or the root should have bound F to the tunnel. In addition, it should have also determined if the tunnel is a new one or an existing one which this M-Flow spec is bound with.

In the former case, a new QPI is created for the new tunnel and bound to F. In the latter case, F is bound to the QPI of the existing tunnel.

3.1.2. In-Band Tunnel Signaling at Transit LSR

As a transit LSR receives a P2MP tunnel signaling message for M-Flow spec F, it does the following:

Case 1: bind_candidate(F) gives a candidate tunnel T, and T is then bound to F. Therefore the LSR becomes a branching LSR. It does the following:

- a. Merge F into T.mflow_specs with
mflow_specs_merge(T, F)
- b. The branching LSR MPLS signaling procedure of specific signaling protocol is then performed

Case 2: Otherwise, it is still an unknown tunnel. It does:

- a. Merge F to T.mflow_specs using

```
mflow_specs_merge(T, F);
```

- b. The non-branching LSR MPLS signaling procedure of the specific protocol is then performed.

3.1.3. In-Band Tunnel Signaling at Root LER

Assume an M-Flow spec F is received at the root mPMBR from MPLS backbone.

- Case 1: F is an M-Flow spec already bound to a tunnel ingress. Therefore the corresponding QPI of the tunnel is used for F.
- Case 2: F is not bound to an existing tunnel yet but F's binding policy is compatible with an M-Flow spec which has been bound to a P2MP tunnel. Therefore, F can then be bound to the same tunnel, and its QPI is used for F.
- Case 3: Neither Case 1 or Case 2 can bind the M-Flow spec.
 - a. Determine if a new tunnel or an existing tunnel is used for the M-Flow spec, based on binding policy and other requirements; if it is a new tunnel, complete the rest of tunnel signaling using the signaling protocol;
 - b. Create a QPI for the tunnel and bound it to F;
 - c. The new QPI is added to root mPMBR PIM as an quasi-PIM oif, and F is mapped to PIM's downstream state;
 - d. Complete the rest signaling at root with specific tunnel signaling procedures

3.1.4. Considerations for Load Balance and Traffic Engineering(TE)

Each LSR including any transit node in the MPLS backbone uses the combination of aggregation and load-balance policies, as well as traffic engineering requirements to decide if an existing tunnel should be shared for an M-Flow spec, and how to route it if the M-Flow spec is to use a separate tunnel.

For example, a transit LSR may decide to merge a new M-Flow spec into an existing P2MP tunnel to avoid allocating new network resources it; or decides to reserve resources initially to be ready for a new tunnel, but once an upstream LSR decides to use an existing tunnel for the M-Flow spec, it will release the resources it has reserved

earlier. But if eventually a new tunnel is created, the reserved resources will be used by the new tunnel.

3.2. Operational Procedures for PIM RPT to SPT switch

This framework uses mPMBR PIM's native RPT to SPT switch to initiate the corresponding traffic switch from the RPT's P2MP tunnel to the SPT's P2MP tunnel. At the downstream egress mPMBR, when a (S, G, RPT) state is created for the bound QPI, the mPMBR's PIM-MPLS inter-working module adds the corresponding M-Flow spec F of Type-4 to the tunnel's M-Flow specs, using `mflow_specs_merge(T, F)`. The new M-Flow spec is then sent to the root mPMBR.

When F is received by the root mPMBR, the native PIM procedure will be performed at the mPMBR to complete the switch. This procedure determines if (S, G) traffic should be stopped on the RPT as traffic now is being received from the SPT.

4. OAM & P

This section is to be completed in future.

5 Security Considerations

There is no additional security requirement for this work.

6 IANA Considerations

There is no IANA impact from the framework.

7 References

7.1 Normative References

[PIM] B. Fenner, M. Handley, H. Holbrook, I. Kouvelas, "Protocol Independent Multicast - Sparse Mode (PIM-SM): Protocol Specification (Revised)", RFC 4601, August 2006.

[mRSVP-TE] R., Li, Q. Zhao, and C. Jacquenet, "Receiver-Driven Multicast Traffic Engineered Label Switched Paths", draft-lzj-mpls-receiver-driven-multicast-rsvp-te-00 (work in progress), March 2012.

[mLDP] Wijnands, IJ., Minei, I., Kompella, K., and B. Thomas, "Label Distribution Protocol Extensions for Point-to-Multipoint and Multipoint-to-Multipoint Label Switched Paths", RFC 6388, November 2011.

[RFC4875] R. Aggarwal, D. Papadimitriou, S. Yasukawa, "Extensions to Resource Reservation Protocol - Traffic Engineering (RSVP-TE) for Point-to-Multipoint TE Label Switched Paths (LSPs)", RFC 4875, May 2007

7.2 Informative References

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

[RFC6513] E. Rosen, R. Aggarwal, "Multicast in MPLS/BGP IP VPNs", RFC 6513, Feb. 2012.

[RFC6514] R. Aggarwal, E. Rosen, T. Morin, Y. Rekhter, "BGP Encodings and Procedures for Multicast in MPLS/BGP IP VPNs",

RFC 6514, Feb. 2012.

[RFC6517] T. Morin, B. Niven-Jenkins, Y. Kamite, R. Zhang, N. Leymann, N. Bitar, "Mandatory Features in a Layer 3 Multicast BGP/MPLS VPN Solution", RFC 6517, Feb., 2012

[RFC6037] E. Rosen, Y. Cai, and IJ. Wijnands, "Cisco Systems' Solution for Multicast in BGP/MPLS IP VPNs", RFC 6037, October 2010.

[I-D.Wijnands-mldp-inband] IJ. Wijnands, Ed., T. Eckert, N. Leymann, M. Napierala, "Multipoint LDP in-band signaling for Point-to-Multipoint and Multipoint-to-Multipoint Label Switched Paths", draft-ietf-mpls-mldp-in-band-signaling-06, December 1, 2011

[I-D.rekhter-pim-sm-over-mldp] Rekhter, Y., Aggarwal, R., and N. Leymann, "Carrying PIM-SM in ASM mode Trees over P2MP mLDP LSPs", draft-rekhter-pim-sm-over-mldp-04, August 2010.

[I-D.lin-mrsvp-te-mvpn] L. Han, "Multicast VPN Support by Receiver-Driven Multicast Extensions to RSVP-TE", draft-hlj-l3vpn-mvpn-mrsvp-te-00, July 2012

Authors' Addresses

Bisong Tao
2330 Central Expressway
Santa Clara, CA 95050
EMail: roberttao@huawei.com

Others(TBD)

Acknowledgement

The author(s) would like to thank the members of Huawei USA IP/MPLS Team for their helpful review comments during the work of this draft.

MPLS Working Group
Internet-Draft
Intended status: Standards Track
Expires: January 10, 2013

A. Atlas
R. Torvi
M. Jork
Juniper Networks
July 9, 2012

Ingress Protection for RSVP-TE p2p and p2mp LSPs
draft-torvi-mpls-rsvp-ingress-protection-00

Abstract

Protection against node failure is important for RSVP-TE LSPs, whether point-to-point or point-to-multipoint. While [RFC4090] provides a mechanism for node protection, it does not specify how to protect against failure of the ingress node. This document specifies the RSVP extensions to support ingress node protection and describes the necessary processing behavior.

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 10, 2013.

Copyright Notice

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

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of

the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction	3
2. Failure Detection Issues and Solution	4
3. Description of Behavior	5
3.1. Ingress Node	5
3.1.1. Required Configuration Information	5
3.1.2. Signaling Behavior	6
3.2. Backup Node	7
3.2.1. Behavior for On-Forwarding-Path Backup Node	7
3.2.2. Behavior for Off-Forwarding-Path Backup Node	7
3.3. Merge Node	8
3.4. Global Repair	8
3.5. Ingress Revival and Administrative Switching	9
4. RSVP Extensions	9
4.1. INGRESS-PROTECTION object	9
5. Normative References	10
Authors' Addresses	10

1. Introduction

It is desirable to protect RSVP-TE LSPs, whether p2p or p2mp, against ingress failure. To do this, a backup node must be pre-identified and prepared with the necessary state so that it can forward traffic when necessary.

Conceptually, a proxy ingress node is created that starts the RSVP signaling. The explicit path of the LSP goes from the proxy ingress node to the backup node and then to the real ingress node. The behavior and signaling for the proxy ingress node is done by the real ingress node.

The backup node must be only one logical hop away from the ingress, whether that be via a direct link or a tunnel.

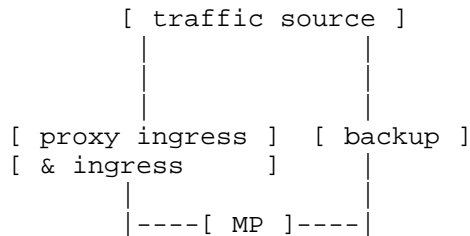


Figure 1: Example Protected LSP with Proxy Ingress Node

There are three different scenarios that this document addresses for ingress protection. All three can be handled using the same set of signaling defined in this document.

- A. Traffic Source detects failure The traffic source(s) can rapidly determine that the ingress has failed and switch over to sending traffic to the backup node. When this mode is specified, the backup node will forward any appropriately received traffic along its bypass tunnel to the merge point(s).
- B. MP detects failure The traffic source(s) always send traffic to both the ingress and backup nodes. The backup node always forwards traffic along its bypass tunnel to the merge point(s). Each MP determines whether the ingress node has failed and, if so, switches over to accepting the traffic from the backup node.

C. Backup detects failure The traffic source(s) always send traffic to both the ingress and backup nodes. The backup node does not forward the received traffic from the traffic source under normal conditions. When the backup node determines that the ingress node has failed, the backup node starts forwarding the traffic alongs its bypass tunnel(s) to the merge point(s).

For all three scenarios, it is necessary for the backup node to know the merge point(s) and associated MPLS labels. This is accomplished by having the RSVP Path and RESV messages go through the backup node, although the forwarding path need not go through the backup node. There are two cases of interest - on-forwarding-path and off-forwarding-path. In the on-forwarding-path case, the backup node is already the immediate node after the ingress node for the LSP. In the off-forwarding-path, the backup node is not the immediate node after the ingress node for all asociated sub-LSPs.

For ingress protection to be functional, the backup node must have access and knowledge of the appropriate traffic to send into the protected LSP. The ingress node must be capable of describing the traffic to the backup node.

Once the backup node has the necessary state for the LSP, including the set of merge points, the backup node can use bypass tunnels as described in [RFC4090]. If the LSP is a point-to-multipoint, then the backup node has the option of choosing to use a bypass p2mp tunnel for protection.

Finally, an assumption of local protection is that a global repair mechanism will occur to replace the patched LSP with a new fully functional one. To do this, it is necessary that the backup node be able to enact a global repair that still allows sharing of bandwidth resources between the old and new LSPs.

2. Failure Detection Issues and Solution

For each of the different scenarios, the details of detecting an ingress node failure can vary. This document does not specify the details of how to do so, but it is possible using either appropriately routed BFD sessions or direct link information.

For traffic-source detection and fail-over, the traffic source can merely monitor the state of the direct links over which traffic is sent to the ingress.

For MP detection, the MP can be configured with the appropriate BFD discriminators used on the BFD sessions. It is desirable for the MP

to know that the ingress can't send traffic to the MP or downstream (for when the ingress is protecting against the MP failure). The appropriate BFD discriminators will vary by MP; they are not signaled in the RSVP extensions described in this draft.

For backup node detection, the backup node can be configured with the appropriate BFD discriminators used on the BFD sessions. Again, they are not signaled in the RSVP extensions described in this draft.

3. Description of Behavior

3.1. Ingress Node

3.1.1. Required Configuration Information

The ingress node must be configured with four pieces of information for these extensions to work.

Backup Node Address The ingress node must know an IP address for the backup node that can be included in the ERO.

Protection Scenario The ingress node must know whether the traffic source, backup node, or merge point(s) will be responsible for handling fail-over.

Ingress-Protector-Context-Id The Ingress-Protector-Context-Id is used for the Extended Session ID in ingress-protected LSPs instead of using the ingress node's loopback address. The Ingress-Protector-Context-Id should not be the same as another address associated with a router that may signal TE LSPs. By having an Ingress-Protector-Context-Id, the backup node can perform global repair.

Application Traffic Identifier The ingress and backup node must both know what application traffic should be directed into the LSP. A commonly understood Application Traffic Identifier is sent between the ingress and backup nodes in RSVP signaling. The exact meaning of the identifier should be configured similarly at both the ingress and backup nodes. The Application Traffic Identifier is understood within the unique context of the Ingress-Protector-Context-Id.

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

3.1.2. Signaling Behavior

The ingress node 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 node is not the first node on the path (for all sub-LSPs), then insert at the beginning of the ERO first the backup node and then the ingress node.
3. Change the IPv4 tunnel sender address in the Sender Template Object to be that of the Ingress-Protector-Context-Id.
4. In the Path RRO, instead of recording the ingress node's address, replace it with the Ingress-Protector-Context-Id.
5. Leave the HOP object populated as usual with information for the ingress-node.
6. Add the INGRESS-PROTECTION object to the Path message. Allocate a second LSP-ID to be used in the INGRESS-PROTECTION object.
7. The RSVP Path message is sent to the backup node as normal. Since the backup node must be only one logical hop away from the ingress, normal RSVP signaling can be used.

When the backup node is off the forwarding path, there are additional behaviors for the ingress node to do when it is handling the associated PATH and RESV messages.

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 node, the ingress node SHOULD check the Failure Scenario specified in the INGRESS-PROTECTION object and, if it is not the "MP detects failure" scenario, then 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 node or from a different node. The stored associated PATH message contains an INGRESS-PROTECTION object that identifies the backup 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 node is on the forwarding path, then a RESV is received with an INGRESS-PROTECTION object and an NHOP that matches the backup node. In this case, the ingress node's address will not appear after the backup node in the RRO. The ingress node should set up ingress forwarding state, just as is done if the LSP weren't ingress-node protected.

3.2. Backup Node

An LER determines that the LSP is ingress-protected based upon the presence of the INGRESS-PROTECTION object in the PATH message. An LER can further determine that it is the backup node if one of its addresses is listed as the backup node in the INGRESS-PROTECTION object.

3.2.1. Behavior for On-Forwarding-Path Backup Node

If the backup node is on the forwarding path, then the backup node MUST remove the INGRESS-PROTECTION object from the PATH message before forwarding it.

If the failure scenario is either "MP-detected" or "Backup-detected", then the backup node is responsible for determining if the ingress node has failed and forwarding the identified traffic from the traffic source(s) to the next-hop(s) on the LSP instead of forwarding the traffic from the ingress node.

When the backup node receives a RESV message, it should add back in the INGRESS-PROTECTION object before forwarding it.

3.2.2. Behavior for Off-Forwarding-Path Backup Node

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

When the backup node receives a RESV message with the INGRESS-PROTECTION object, the backup node records an IMPLICIT-NULL label in the RRO. The backup node creates the appropriate forwarding state for the failure scenario specified. For the "MP-detected" and "traffic-source-detected", this means that backup node forwards any received identified traffic into the bypass tunnel(s) to the merge

point(s). For the "backup-detected", this means that the backup node creates state to quickly determine the ingress has failed and switch to sending any received identified traffic into the bypass tunnel(s) to the merge point(s). Then the backup node forwards the RESV message to the ingress node, which is acting for the proxy ingress.

If the backup node doesn't have a bypass tunnel to a merge point, then the backup node can wait to send the RESV until such has been created or it can send a Path Err with an Error Code of "Routing Problem (24)" and a new Error Value sub-code of "No Bypass Tunnel to Merge Point (TBD)".

3.3. Merge Node

An LSR that is serving as a Merge Node may need to support the INGRESS-PROTECTION object and functionality defined in this specification if the LSP is ingress-protected where the failure scenario is "MP-detected". An LSR can determine that it must be a merge point by examining the INGRESS-PROTECTION object and determining that it is neither the ingress node nor the backup node and the PHOP is the ingress node.

In that case, when the LSR receives a PATH message with an INGRESS-PROTECTION object, the LSR MUST remove the INGRESS-PROTECTION object before forwarding on the PATH message.

If the failure scenario specified is "MP-detected", the MP must connect up the fast-failure detection (as configured) to accepting backup traffic received from the backup node. There are a number of different ways that the MP can enforce not forwarding traffic normally received from the backup node. For instance, first, any LSPs set up from the backup node should not be signaled with an IMPLICIT NULL label and second, the associated label for the ingress-protected LSP could be set to normally discard inside that context.

When the MP receives a RESV message whose matching PATH state had an INGRESS-PROTECTION object, the MP SHOULD add the INGRESS-PROTECTION object to the RESV message before forwarding it.

3.4. Global Repair

When the backup node learns the ingress node has failed (e.g. via the IGP), then the backup node can compute new ERO(s) and signal the new LSP so that it no longer relies upon local repair. To do this, the backup node uses the same Ingress-Protector-Context-Id as the Ipv4 tunnel sender address in the Sender Template Object and uses the previously allocated second LSP-ID signaled in the INGRESS-PROTECTION object. This allows the new LSP to share resources with the old LSP.

3.5. Ingress Revival and Administrative Switching

In a future version, it is intended to describe the behavior when the ingress node comes back and how to handle management-triggered switches from ingress to backup node and vice versa.

4. RSVP Extensions

4.1. INGRESS-PROTECTION object

Class-Num = TBD

C-Type = TBD

0	1	2	3
Length (bytes)	Class-Num	C-Type	
Backup Node Address			
Ingress Node Address			
Application Traffic Identifier			
Secondary LSP ID	Protection Scenario	Flags	

Figure 2: INGRESS-PROTECTION object

Backup Node Address

Ingress Node Address

Application Traffic Identifier

Ingress-Protector-Context-Id

Secondary LSP ID

Protection Scenario Indicates if (1) traffic source(s), (2) backup node, (3) or merge point(s) will handle the fail-over.

Control Flags Backup sent flags: (0x01)Ingress-Protection in-use,
(0x02)Ingress Detected Down, (0x04)Admin Override caused Ingress-
Protection-in-use. Ingress sent flags: (0x08)Revert Control to
Ingress, (0x10)Force control to Backup

5. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC2205] Braden, B., Zhang, L., Berson, S., Herzog, S., and S. Jamin, "Resource ReSerVation Protocol (RSVP) -- Version 1 Functional Specification", RFC 2205, September 1997.
- [RFC3209] Awduche, D., Berger, L., Gan, D., Li, T., Srinivasan, V., and G. Swallow, "RSVP-TE: Extensions to RSVP for LSP Tunnels", RFC 3209, December 2001.
- [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.

Authors' Addresses

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

Email: akatlas@juniper.net

Raveendra Torvi
Juniper Networks
10 Technology Park Drive
Westford, MA 01886
USA

Email: rtorvi@juniper.net

Markus Jork
Juniper Networks
10 Technology Park Drive
Westford, MA 01886
USA

Email: mjork@juniper.net

Network Working Group
Internet-Draft
Intended status: Informational
Expires: January 7, 2013

Y. Weingarten

S. Aldrin
Huawei Technologies
July 6, 2012

Requirements for MPLS Shared Mesh Protection
draft-weingarten-mpls-smp-requirements-00.txt

Abstract

This document presents the basic network objectives for the behavior of shared mesh protection (SMP) not based on control-plane support. This is an expansion of the basic requirements presented in the MPLS Transport Profile Requirements (RFC5654) and MPLS Transport Profile Survivability Framework (RFC6372), documents. This document should be used as a basis for the definition of the mechanism that would be used to implement SMP for MPLS data paths, in networks that do not employ a control plane for its operation.

Status of this Memo

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

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

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

This Internet-Draft will expire on January 7, 2013.

Copyright Notice

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

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

to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction	3
1.1. Protection or Restoration	4
1.2. Scope of document	4
1.2.1. Relationship to MPLS-TP	5
1.3. Contributing Authors	5
2. Terminology and Notation	5
2.1. Acronyms	5
3. SMP Architecture	5
3.1. Coordination of resources	7
4. SMP Network Objectives	7
4.1. Configuration and resource reservation	7
4.1.1. Querying resource availability	7
4.2. Control plane or data plane	8
4.3. Multiple faults	8
4.4. Notification	9
4.5. Protection switching time	9
4.6. Timers	9
5. IANA Considerations	9
6. Managability Considerations	10
7. Security Considerations	10
8. Acknowledgements	10
9. Normative References	10
Authors' Addresses	10

1. Introduction

MPLS transport networks can be characterized as being a network of connections between nodes within a mesh of nodes and the links between them. The connections, that may be between neighboring nodes, i.e. spanning a single physical link, or spanning a path of several nodes, constitute the Label Switched Paths (LSP) that transport packets between the endpoints of these paths. The survivability of these connections, as described in [RFC6372], is a critical aspect for various service providers that are bound by Service Level Agreements (SLA) with their customers.

MPLS provides control-plane tools to support various survivability schemes (Editor's note - add references). In addition, recent efforts in the IETF have started providing for data-plane tools to address aspects of data protection. In particular, [RFC6378] defines a set of triggers and coordination protocol for 1:1 and 1+1 linear protection of p2p paths.

When considering a full-mesh network and the protection of different paths that criss-cross the mesh, it is possible to conserve the amount of protection resources needed to protect the different data paths. As pointed out in [RFC6372] and [RFC4428], applying 1+1 linear protection, requires that resources are allocated and used by both the working and protection paths. Applying 1:1 protection requires that all of the resources are allocated, but allows the resources of the protection path to be utilized for pre-emptible extra traffic. Extending this to 1:n or m:n protection allows the resources of the protection path to be shared in the protection of several working paths. However, there is a limitation in 1:n protection architectures - that all of the n+1 paths must have identical endpoints.

Shared Mesh Protection (SMP) supports a limited form of resource sharing of the protection resources, while providing protection for multiple data paths that may not have common endpoints and do not share common points of failure. The basic configuration for data paths that employ SMP is shown in Figure 1. In this figure, we show two working paths [ABCDE] and [VWXYZ] that are protected (by 1:1 linear protection) protection paths [APQRE] and [VPQRZ] respectively. The segment [PQR] and all of its protection resources are shared by both of the protection paths.

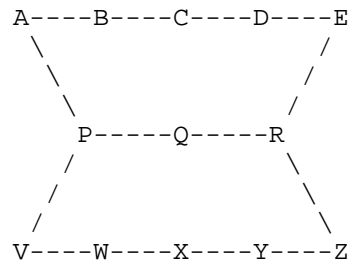


Figure 1: Basic SMP architecture

1.1. Protection or Restoration

[RFC6372], based upon the definitions in [RFC4427] makes the differentiation between "protection" and "restoration" dependent upon the dynamism of the resource allocation. In SMP, the resources of the protection paths are reserved at the time of path creation. However, the allocation of the full resources, at least for the shared segments will only be finalized at the time that the protection path is actually activated. Therefore, for the purists - regarding the terminology - SMP lies somewhere between protection and restoration.

1.2. Scope of document

[RFC5654] also establishes that MPLS-TP should support shared protection (Requirement 68) and that MPLS-TP must support sharing of protection resources (Requirement 69). This document presents the network objectives and a framework for applying SMP within an MPLS network, without the use of control-plane protocols. There are existing control-plane solutions for SMP within MPLS, however we address those networks that for some reason, e.g. service provider preferences or limitations, do not employ a full control plane operation, or require service restoration faster than achievable with control plane mechanisms.

The network objectives will also address possible additional restrictions of the behavior of SMP in statically configured operator networks.

Protection Switching Control (PSC) defines the protection switching process, logic and protocol messages, based on the SMP actions sent and received from participating LSRs. Definition of logic and specific protocol messaging is out of scope of this document.

1.2.1. Relationship to MPLS-TP

While some of the restrictions presented by this framework originate from the considerations of transport networks, there is no real constraint of the information presented here being applied to general MPLS networks, and not necessarily as part of the Transport Profile of MPLS.

1.3. Contributing Authors

David Allan, Gregory Mirsky, Daniel King

2. Terminology and 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].

The terminology used in this document is based on the terminology defined in the MPLS-TP Survivability Framework document [RFC6372] which in-turn is based on [RFC4427].

2.1. Acronyms

This draft uses the following acronyms:

LSP Label Switched Path
PSC Protection State Coordination protocol
SLA Service Level Agreement
SMP Shared Mesh Protection
SRLG Shared Risk Link Group

3. SMP Architecture

Figure 1 shows a very basic configuration of working and protection paths that may employ SMP. We may consider a slightly more involved configuration, such as the one in Figure 2 in order to identify certain basic characteristics of an SMP mesh network.

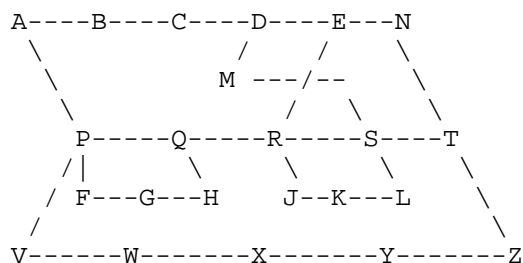


Figure 2: Larger sample SMP architecture

Consider the network presented in Figure 2. There are five working paths - [ABCDE], [MDEN], [FGH], [JKL], and [VWXYZ]. Each of these has a corresponding protection path - [APQRE] (p1), [MSTN] (p2), [FPQH] (p3), [JRSL] (p4), and [VPQRSTZ] (p5). The following segments are shared by two or more of the protection paths - [PQ] is shared by p1, p3, and p5, [QR] is shared by p1 and p5, [RS] is shared by p4 and p5, and [ST] is shared by p2 and p5. In addition, we assume that the available protection resources for these shared segments are not sufficient to support the complete traffic capacity of the respective working paths that may use the protection paths. We can further observe that the main feature of the network that defines it as an SMP network is the fact that the segment [PQRST], that is a sub-segment of p5, is the union of all the shared segments while being a whole shared segment of one of the protection paths.

In other words, the main feature of an SMP "protection domain" will be the segment that is the union of all the shared segments of the protection paths. We can further identify "protection groups" as the different protection paths that share a common segment. For example, referring to Figure 2, we have the following protection groups - {p1, p3, p5} for [PQ], {p1, p5} for [QR], {p4, p5} for [RS], {p2, p5} for [ST].

Typical deployment of SMP would require various network planning activities. These would include:

- o Identification of key services that require protection, and determining the number of working and protection paths.
- o Reviewing network topology to determine which working or protection paths are required to be disjointed from each other, and exclude specified resources such as links, nodes, shared risk link groups (SRLGs).

3.1. Coordination of resources

When a protection switch is triggered by any fault condition or operator command, the SMP network must perform two operations almost simultaneously - switch data traffic over to a protection path and verify that the shared resources are allocated for this protection path. The allocation of resources is dependent upon their availability at each of the shared segments.

When the reserved resources of the shared segments are allocated for a particular protection path, there may not be sufficient resources available for an additional protection path. This then implies that if an additional working path triggers a protection switch the allocation of the resources may fail and MUST be treated as described below in Section 4.3. In order to optimize the operation of the allocation and preparing for cases of multiple working path failures, the allocation of the shared resources SHALL be coordinated between the different working paths in the SMP network.

4. SMP Network Objectives

4.1. Configuration and resource reservation

SMP is a survivability mechanism that is based on pre-configuration of the network working paths and the corresponding protection paths. This configuration may be based on either a control protocol or static configuration by the management system. The protection relationship between the working and protection paths SHOULD be configured and the shared segments of the protection path must be identified prior to use of the protection paths.

As opposed to the case of simple linear protection, where the relationship between the working and protection paths is defined, the resources for the protection path may be fully committed for the unshared portions of the protection path. The protection path in the case of SMP consists of segments that are dedicated to the protection of the related working path and also segments that are shared with other protection paths. On the shared segments, the protection resources may be reserved but would not be allocated until requested as part of a protection switch.

4.1.1. Querying resource availability

When a working path identifies a protection switching trigger it SHOULD verify that the necessary protection resources are available on the protection path. The resources may not be available because they have been allocated to the protection of a higher priority

working path, as described above.

4.2. Control plane or data plane

As stated in both [RFC6372] and [RFC4428] full control of SMP, including both configuration and the coordination of the protection switching is potentially very complex. Therefore, it is suggested that this be carried out under the control of a dynamic control plane similar to GMPLS [RFC3945]. In fact, implementations for SMP with GMPLS exist and the general principles of its operation are well known, if not fully documented.

There are, however, operators, in particular in the transport sector, that do not operate their MPLS networks under the control of a control plane and require the ability of performing SMP protection while utilizing data-plane tools for coordination of the protection switching. This requirement is emphasized in different areas of [RFC5654] for MPLS-TP environments. Therefore, it is imperative that it be possible to perform all of the coordination needed for SMP via data plane operations.

4.3. Multiple faults

If more than one working path is triggering a protection switch there are different possible actions that the SMP network may apply. The basic MPLS action MAY allow all of the protection paths to share the resources of the shared segments, for those networks that support multiplexing packets over the shared segments. For those networks, in particular for networks that support the requirements in [RFC5654] [and in particular support for requirement 58], that require the exclusive use of the protection resources, the following behavior SHOULD be supported:

- o Relative priority MAY be assigned to each of the working paths that share a common protection segment
- o Resources of the shared segments SHALL be allocated to the protection path according to the highest priority amongst those requesting use of the resources.
- o If the protection resources are currently in use by a protection path, whose working path has a lower priority, resources SHALL be allocated to the path with higher priority. Traffic with lower priority MAY use available resources or MAY be interrupted.
- o Shared segment resources MAY be used by existing traffic and higher priority traffic for a short period until preemption is completed.

4.4. Notification

When a working path identifies a trigger for implementing a switchover to the protection path, it SHALL attempt to switchover the traffic to the protection path and requesting the allocation of the resources for this protected traffic. If the protection path is not able to allocate the necessary resources (e.g. the resources are being used for protected traffic of higher priority), a notification SHALL be sent to both endpoints of the requesting working path indicating that the requested switchover cannot be fulfilled.

Similarly, if preemption is supported and as a result of the allocation of resources to a different working path that triggered a protection switch, the resources currently allocated for a particular working path are being preempted then a notification SHALL be sent to the endpoints of the working path whose traffic is being preempted indicating that the resources are being preempted.

4.5. Protection switching time

In general, protection switching time is defined as the interval after a switching trigger is identified until the traffic begins to be transmitted on the protection path. This time is exclusive of the time needed to complete preemption of existing traffic on the shared segments as described in Section 4.3. Therefore, support for a switching time of 50ms is dependent upon the initial switchover to the protection path

4.6. Timers

In order to prevent multiple switching actions for a single switching trigger, SMP SHOULD be controlled a hold-off timer that would allow lower level mechanisms to complete their switching actions before invoking SMP protection actions.

In addition, to prevent an unstable recovering working path from invoking intermittent switching operation, SMP SHOULD employ a wait-to-restore timer during any reversion switching.

5. IANA Considerations

This document makes no request of IANA.

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

6. Managability Considerations

To be added in future version.

7. Security Considerations

To be added in future version.

8. Acknowledgements

TBD

9. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC5654] Niven-Jenkins, B., Nadeau, T., and C. Pignataro, "Requirements for the Transport Profile of MPLS", RFC 5654, Sept 2009.
- [RFC6372] Sprecher, N. and A. Farrel, "MPLS-TP Survivability Framework", RFC 6372, Sept 2011.
- [RFC6378] Sprecher, N., Bryant, S., Osborne, E., Fulignoli, A., and Y. Weingarten, "MPLS-TP Linear Protection", RFC 6378, Nov 2011.
- [RFC3945] Mannie, E., "Generalized Multi-Protocol Label Switching (GMPLS) Architecture", RFC 3945, Oct 2004.
- [RFC4427] Mannie, E. and D. Papadimitriou, "Recovery (Protection and Restoration) Terminology for GMPLS", RFC 4427, March 2006.
- [RFC4428] Mannie, E. and D. Papadimitriou, "Analysis of Generalized Multi-Protocol Label Switching (GMPLS)-based Recovery Mechanisms (including Protection and Restoration)", RFC 4428, March 2006.

Authors' Addresses

Yaacov Weingarten
34 Hagefen St.
Karnei Shomron, 4485500
Israel

Phone:
Email: wyaacov@gmail.com

Sam Aldrin
Huawei Technologies
2330 Central Express Way
Santa Clara, CA 95951
United States

Email: aldrin.ietf@gmail.com

Network working group
Internet Draft
Category: Standard Track

X. Xu
Huawei
M. Eubanks
AmericaFree.TV
L. Yong
Z. Li
Huawei
N. Sheth
Juniper
Y. Fan
China Telecom

Expires: January 2013

July 13, 2012

Encapsulating MPLS in UDP

draft-xu-mpls-in-udp-02

Status of this Memo

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

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

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

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

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

This Internet-Draft will expire on January 13, 2013.

Copyright Notice

Copyright (c) 2009 IETF Trust and the persons identified as the document authors. All rights reserved.

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

Abstract

This document specifies one additional IP-based encapsulation technology for MPLS packets referred to as MPLS-in-UDP, which is intended to facilitate load-balancing the traffic of various MPLS applications such as MPLS-based L2VPN and L3VPN in the core of IP-enabled packet switch networks.

Conventions used in this document

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

Table of Contents

1. Introduction	3
2. Terminology	3
3. Encapsulation in UDP.....	4
4. Signaling for Encapsulation in UDP	5
5. Processing Functions.....	5
6. Applicability	6
7. Security Considerations	6
8. IANA Considerations	6
9. Acknowledgements	6
10. References	6
10.1. Normative References	6
10.2. Informative References	6
Authors' Addresses	7

1. Introduction

Equal Cost Multi-Path (ECMP) and Link Aggregation Group (LAG) are widely used in the core of IP-enabled Packet Switch Networks (PSN) for load-balancing purposes. Most core routers (i.e., P routers) in the IP-enabled PSN are capable of load-balancing IP traffic flows across ECMP paths and/or LAG based on the hash of the five-tuple of UDP/TCP packets (i.e., source IP address, destination IP address, source port, destination port, and protocol) or some fields in the IP header of non-UDP/TCP packets (e.g., source IP address, destination IP address). However, with existing IP-based encapsulation methods as defined in [RFC4023] (e.g., MPLS-in-IP and MPLS-in-GRE), distinct customer traffic flows of various MPLS applications (e.g., MPLS-based L2VPN or L3VPN) between a given PE pair would be encapsulated with the same IP or GRE tunnel header prior to traversing the IP core. Since the encapsulating traffic is neither TCP nor UDP traffic, core routers could only perform hash calculation on the fields in the IP header of IP or GRE tunnels. As a result, core routers could not achieve an effective load-balancing for these traffic flows in the network due to the lack of adequate entropy information.

[RFC5640] describes a method for improving the load-balancing in Software mesh networks [RFC5565]. However, this method requires core routers to be able to perform hash calculation on the fields including the "load-balancing" field contained in the L2TPv3 or GRE tunnel header. [Entropy-Label] proposes to use the "entropy labels" for achieving a better load-balancing for MPLS traffic flows in the core of MPLS-enabled PSN. Although the entropy label could be inserted in the "Key" field of the GRE header by ingress PE routers in the case where the PSN is IP enabled rather than MPLS enabled, it still requires core routers to be capable of performing hash calculation on the "entropy label" contained in the GRE tunnel header. Any of the above load-balancing methods requires a change to the data plane of core routers.

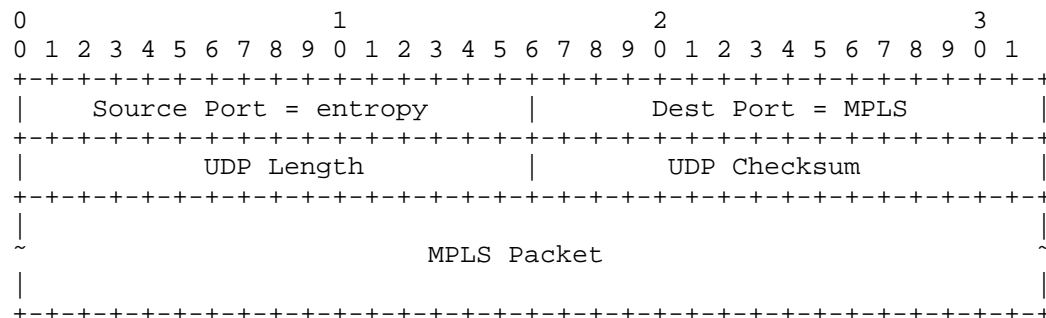
This document describes a new IP-based encapsulation method for MPLS packets referred to as MPLS-in-UDP, which is intended to facilitate load-balancing the traffic of various MPLS applications such as MPLS-based L2VPN and L3VPN in the core of IP-enabled packet switch networks where the core routers could not be upgraded due to some reason.

2. Terminology

This memo makes use of the terms defined in [RFC4364] and [RFC4664].

3. Encapsulation in UDP

MPLS-in-IP messages have the following format:



Source Port of UDP

This field contains an entropy value that is generated by the ingress PE router. For example, the entropy value can be generated by performing hash calculation on certain fields in the customer packets (e.g., the five tuple of UDP/TCP packets). To ensure that the source port number is always in the range 49152 to 65535 which is required in some cases, instead of calculating a 16-bit hash, the ingress PE router would calculate a 14-bit hash and use those 14 bits as the least significant bits of the source port field while the most significant two bits would be set to binary 11. That would still convey 14 bits of entropy information which is also enough in practice.

Destination Port of UDP

This field is set to a value (TBD) indicating the MPLS packet encapsulated in the UDP header is a MPLS unicast one or a MPLS multicast one.

UDP Length

The usage of this field is in accordance with the current UDP specification.

UDP Checksum

The usage of this field is in accordance with the current UDP specification. To simplify the operation on

egress PE router, this field is recommended to be set to zero.

4. Signaling for Encapsulation in UDP

PE routers could signal the UDP tunnel encapsulation information among them by some means.

In the case when BGP is used in the MPLS applications (e.g., BGP/MPLS IP VPN [RFC4364]), the MPLS-in-UDP encapsulation information can be signaled by using the mechanism defined in [RFC 5512]. In this case, a new Tunnel Type code for UDP tunnel technology needs to be assigned by IANA. If there is no explicit encapsulation information to signal using the Encapsulation SAFI for the UDP tunneling protocol, a BGP Encapsulation Extended Community with the Tunnel Type set to the value indicating UDP tunneling protocol would be enough. For example, such extended community could be attached to the update messages for NLRI announcement in the BGP/MPLS IP VPN case, or be attached to the update messages dedicated for auto-discovery in the VPLS [RFC4761, RFC4762] case where BGP-based auto-discovery is used. Otherwise, if more detailed information about the UDP tunnel technology is needed for signaling (e.g., to specify what MPLS application is allowed to use this MPLS-in-UDP encapsulation), a new TLV and even a set of sub-TLVs dedicated for UDP tunnel encapsulation technology that would be contained in the Tunnel Encapsulation attribute needs to be defined.

More details about how to signal the MPLS-in-UDP encapsulation information will be described in a separate document.

5. Processing Functions

This MPLS-in-UDP encapsulation causes MPLS packets to be forwarded through "IP UDP tunnels". When performing MPLS-in-UDP encapsulation by an ingress PE router, the entropy value would be generated by the ingress PE router and then be filled in the Source Port field of the UDP header.

P routers, upon receiving these UDP encapsulated packets, could balance these packets based on the hash of the five-tuple of UDP packets.

Upon receiving these UDP encapsulated packets, egress PE routers would decapsulate them by removing the UDP headers and then process them accordingly.

6. Applicability

Besides the MPLS-based L3VPN [RFC4364] and L2VPN [RFC4761, RFC4762] [E-VPN] applications, MPLS-in-UDP encapsulation could also be used in other MPLS applications including but not limited to 6PE [RFC4798] and PWE3 services.

7. Security Considerations

TBD.

8. IANA Considerations

Two distinct UDP destination port numbers indicating MPLS unicast and MPLS multicast respectively need to be assigned by IANA.

9. Acknowledgements

Thanks to Shane Amante, Dino Farinacci, Keshava A K, Ivan Pepelnjak, Weiguo Hao, Zhenxiao Liu and Xing Tong for their valuable comments on the idea of MPLS-in-UDP encapsulation.

10. References

10.1. Normative References

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

10.2. Informative References

[RFC4364] Rosen, E and Y. Rekhter, "BGP/MPLS IP Virtual Private Networks (VPNs)", RFC 4364, February 2006.

[RFC4664] Andersson, L. and Rosen, E. (Editors), "Framework for Layer 2 Virtual Private Networks (L2VPNs)", RFC 4664, Sept 2006.

[RFC4023] Worster, T., Rekhter, Y., and E. Rosen, "Encapsulating MPLS in IP or GRE", RFC4023, March 2005.

[RFC5640] Filsfils, C., Mohapatra, P., and C. Pignataro, "Load-Balancing for Mesh Softwires", RFC 5640, August 2009.

[RFC6391] Bryant, S., Filsfils, C., Drafz, U., Kompella, V., Regan, J., and S. Amante, "Flow Aware Transport of Pseudowires over an MPLS Packet Switched Network", RFC6391, November 2011

- [Entropy-Label] Kompella, K., Drake, J., Amante, S., Henderickx, W., and L. Yong, "The Use of Entropy Labels in MPLS Forwarding", draft-ietf-mpls-entropy-label-01, work in progress, October, 2011.
- [RFC5512] Mohapatra, P. and E. Rosen, "The BGP Encapsulation Subsequent Address Family Identifier (SAFI) and the BGP Tunnel Encapsulation Attribute", RFC 5512, April 2009.
- [RFC4798] J Declercq et al., "Connecting IPv6 Islands over IPv4 MPLS using IPv6 Provider Edge Routers (6PE)", RFC4798, February 2007.
- [RFC4761] Kompella, K. and Y. Rekhter, "Virtual Private LAN Service (VPLS) Using BGP for Auto-Discovery and Signaling", RFC 4761, January 2007.
- [RFC4762] Lasserre, M. and V. Kompella, "Virtual Private LAN Service (VPLS) Using Label Distribution Protocol (LDP) Signaling", RFC 4762, January 2007.
- [E-VPN] Aggarwal et al., "BGP MPLS Based Ethernet VPN", draft-ietf-12vpn-evpn-00.txt, work in progress, February, 2012.

Authors' Addresses

Xiaohu Xu
Huawei Technologies,
Beijing, China

Phone: +86-10-60610041
Email: xuxiaohu@huawei.com

Marshall Eubanks
AmericaFree.TV LLC
P.O. Box 141
Clifton, Virginia 20124
USA

Phone: +1-703-501-4376
Email: marshall.eubanks@gmail.com

Lucy Yong
Huawei USA

1700 Alma Dr. Suite 500
Plano, TX 75075
US

Email: lucyyong@huawei.com

Nischal Sheth
Juniper Networks
1194 North Mathilda Avenue
Sunnyvale, CA 94089 USA

Email: nsheth@juniper.net

Zhenbin Li
Huawei Technologies,
Beijing, China

Phone: +86-10-60613676
Email: lizhenbin@huawei.com

Yongbing Fan
China Telecom
Guangzhou, China.

Phone: +86 20 38639121
Email: fanyb@gsta.com

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: December 18, 2012

R. Zheng, Ed.
L. Jin, Ed.
ZTE
T. Nadeau, Ed.
Juniper
G. Swallow, Ed.
Cisco
June 16, 2012

Echo Relay Reply mechanism for LSP Ping
draft-zjns-mpls-lsp-ping-relay-reply-00

Abstract

[RFC4379] describes the LSP Ping mechanism to detect data plane failures. In some deployment scenario for the LSP traceroute, a replying LSR may not have the available route to the initiator, and the echo reply message sent to the initiator would be discarded. Thus, the basic idea of traceroute procedure to localize fault could not be achieved. This document describes extensions to LSP Ping mechanism to enable the replying LSR to have the capability to relay the echo reply by a set of routable intermediate nodes to the initiator.

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 18, 2012.

Copyright Notice

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

This document is subject to BCP 78 and the IETF Trust's Legal

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

Table of Contents

1. Introduction	3
1.1. Conventions Used in This Document	3
2. Motivation	3
3. Extensions	4
3.1. Echo Relay Reply message	4
3.2. Relay Node Address Stack	5
4. Procedures	6
4.1. Sending an Echo Request	6
4.2. Receiving an Echo Request	6
4.3. Sending an Echo Relay Reply	7
4.4. Receiving an Echo Relay Reply	8
4.5. Sending an Echo Reply	8
4.6. Receiving an Echo Reply	8
5. Security Considerations	8
6. IANA Considerations	9
6.1. New Message Type	9
6.2. New TLV	9
7. Acknowledgement	9
8. References	9
8.1. Normative References	9
8.2. Informative References	10
Authors' Addresses	10

1. Introduction

This draft describes LSP Ping Echo Relay Reply mechanism that can be used to detect data plane failures in MPLS LSPs that span across multiple domains. A new message referred to as "Echo Relay Reply message" and a new TLV referred to as "Relay Node Address Stack TLV" are defined in this draft.

1.1. Conventions Used in This Document

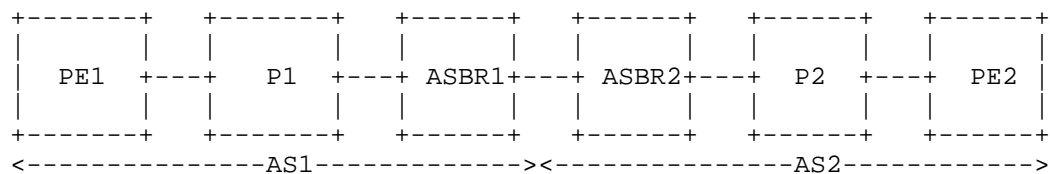
The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in [RFC2119].

2. Motivation

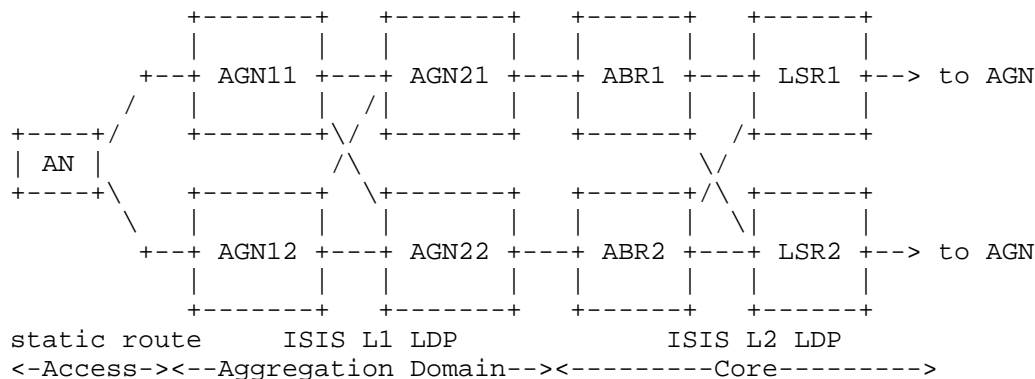
LSP Ping is an efficient OAM mechanism to detect data plane failures and localize faults. The basic LSP Ping mechanism has been described in [RFC4379]. In traceroute mode of LSP Ping procedure, the echo request message is sent to the control plane of each transit LSR, and an echo reply message with proper information are required to send to the initiator at each transit LSR. Then the LSP fault could be localized exactly, and an accurate LSP topology could also be built.

The echo reply would normally be sent back to the initiator via an IPv4/IPv6 UDP packet. The basic requirement is that the replying LSR has reachable IP route to the initiator. However, in some network deployment, the requirement could not be met because of the route control policy.

For inter-AS scenarios, it is common of the providers to NOT distribute the IP addresses of any of the nodes other than the ASBR. If initiating a traceroute procedure on the ingress node PE1 of an LSP from PE1 to PE2, P nodes in the other AS like P2 would be unable to respond to the echo request message for the lack of IP reachable route to PE1.



For the inter-area situation in Seamless MPLS architecture [ietf-mpls-seamless], P nodes in core network would not have IP reachable route to ANs. When tracing an LSP from AN to remote AN, the LSR1/LSR2 node could not make a response to the echo request either, like P2 node in the inter-AS scenario.



This draft describes extensions to LSP Ping mechanism to enable the response from the replying LSR to be relayed back to the initiator. The replying LSR would send the response to a relay node indicated by the Relay Node Address Stack TLV, and the response would be relayed to the next relay node, till to the initiator.

3. Extensions

RFC4379 describes the basic MPLS LSP Ping mechanism, which defines two message types. This draft defines a new message, Echo Relay Reply message. This new message is used to replace Echo Reply message which is sent from the replying LSR to a relay node or from a relay node to another relay node.

In addition, a new TLV named Relay Node Address Stack TLV is defined in this draft, to carry the IP addresses of the possible relay nodes for the replying LSR.

3.1. Echo Relay Reply message

The echo relay reply message is a UDP packet, and the UDP payload has the same format with echo request/reply message. A new message type is requested from IANA.

New Message Type:

Value	Meaning
-------	---------

TBD	MPLS echo relay reply
-----	-----------------------

3.2. Relay Node Address Stack

The Relay Node Address Stack TLV MUST be carried in the echo request, echo reply and echo relay reply messages if the echo reply relayed mechanism described in this draft is required.

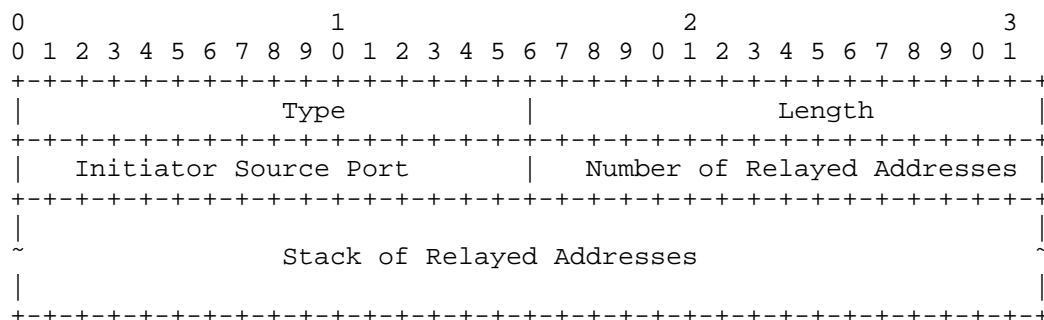
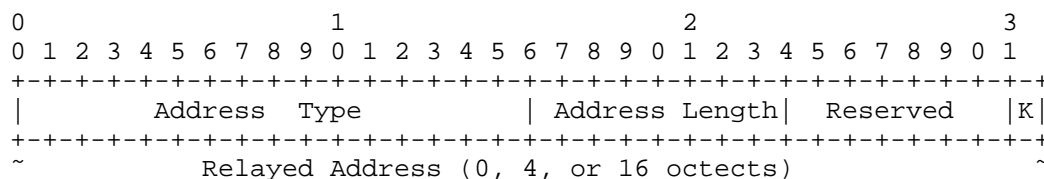


Figure 1: Relay Node Address Stack TLV

- Type: to be assigned by IANA.
- Length: The Length of the Value field in octets.
- Initiator Source Port: The port that the initiator sends the echo request message, and also the port that expected to receive the echo reply message.
- Number of Relayed Addresses: An integer indicating the number of relayed addresses in the stack.
- Stack of Relayed Addresses: A list of relay node addresses.

The format of each relay node address is as below:



[illegible]

Type#	Address Type	Address Length
0	Unspecified	0
1	IPv4	4
2	IPv6	16

Reserved: This field is reserved for future use and MUST be set to zero.

K bit:

If the K bit is set to 1, then this sub-TLV SHOULD be kept in Relay Node Address Stack, SHOULD not be deleted in compress process of section 4.2. The K bit may be set by ASBRs which address would be kept in the stack if necessary.

If the K bit is set to 0, then this sub-TLV SHOULD be processed normally according to section 4.2.

Relayed Address: This field specifies the node address, either IPv4 or IPv6.

4. Procedures

4.1. Sending an Echo Request

The procedures described in Section 4.3 of RFC4379 apply here. In addition, An Relay Node Address Stack TLV MUST be carried in the echo request message.

When the echo request is first sent by initiator, a Relay Node Address Stack TLV with the initiator address in the stack and its source port **MUST** be included.

For the subsequent echo request messages, the initiator would copy the Relay Node Address Stack TLV from the received echo reply message.

4.2. Receiving an Echo Request

In addition to the processes in Section 4.4 of RFC4379, the procedures of the Relay Node Address Stack TLV are defined here.

Upon receiving a Relay Node Address Stack TLV of the echo request

message, the receiver would check the addresses of the stack in sequence from top to bottom, to find out the first public routable IP address. Those address entries behind of the first routable IP address in the address list with K bit set to 0 would be deleted, and the address entry of the replying LSR would be added at the bottom of the stack. Those address entries with K bit set to 1 would be kept in the stack. The updated Relay Node Address Stack TLV would be carried in the response message.

If the replying LSR wishes to hide its routable address information, the address entry added in the stack would be a blank entry with Address Type set to Unspecified. The blank address entry in the receiving echo request would be treated as an unroutable address entry.

If the first routable IP address is the first address in the stack, the replying LSR would respond an echo reply message to the initiator.

If the first routable IP address is of an intermediate node, other than the first address in the stack, the replying LSR would send an echo relay reply instead of an echo reply in response.

4.3. Sending an Echo Relay Reply

The echo relay reply is sent in two cases:

1. When the replying LSR received an echo request with the initiator IP address in the Relay Node Address Stack TLV is IP unroutable, the replying LSR would send an echo relay reply message to the first routable intermediate node. The encapsulation processing of echo relay reply is the same with the procedure of the echo reply described in Section 4.5 of RFC4379, except the destination IP address and the destination UDP port of the message part. The destination IP address of the echo relay reply is set to the first routable IP address from the Relay Node Address Stack TLV, and the destination UDP port is set to 3503.

2. When the intermediate relay node received an echo relay reply with the initiator IP address in the Relay Node Address Stack TLV is IP unroutable, the intermediate relay node would send the echo relay reply to the next relay node with the content of the UDP packet unchanged. The destination IP address of the echo relay reply is set to the first routable IP address from the Relay Node Address Stack TLV. Both the source and destination UDP port should be 3503.

4.4. Receiving an Echo Relay Reply

Upon receiving an echo relay reply message with its address as the destination address in the IP header, the relay node should check the address items in Relay Node Address Stack TLV in sequence and find the first routable node address.

If the first routable address is the top one of the address list, i.e., the initiator address, the relay node should send an echo reply message to the initiator containing the same payload with the echo relay reply message received.

If the first routable address is not the top one of the address list, i.e., another intermediate relay node, the relay node should send an echo relay reply message to this relay node with the payload unchanged.

4.5. Sending an Echo Reply

The echo relay reply is sent in two cases:

1. When the replying LSR received an echo request with the initiator IP address in the Relay Node Address Stack TLV is IP routable, the replying LSR would send an echo reply to the initiator. The processing of echo relay reply is the same with the procedure of the echo reply described in Section 4.5 of RFC4379.

2. When the intermediate relay node LSR received an echo relay reply with the initiator IP address in the Relay Node Address Stack TLV is IP routable, the intermediate relay node would send the echo reply to the initiator with the payload no changes other than the Message Type field. The destination IP address of the echo reply is set to the initiator IP address, and the destination UDP port would be copied from the Initiator Source Port field of the Relay Node Address Stack TLV. The source UDP port should be 3503.

4.6. Receiving an Echo Reply

In addition to the processes in Section 4.6 of RFC4379, the initiator would copy the Relay Node Address Stack TLV received in the echo reply to the next echo request.

5. Security Considerations

In addition to the Security Consideration from [RFC4379], to avoid potential Denial-of-service attack, it is RECOMMENDED that implementations regulate the LSP Ping traffic going to the control

plane. A rate limiter SHOULD be applied to UDP port 3503 of the intermediate node.

The node which acts as a relay node SHOULD validate the relay reply against a set of valid source addresses. An implementation SHOULD provide such filtering capabilities.

If an operator wants to obscure their nodes, then a blank entry may be used in the address stack.

6. IANA Considerations

IANA is requested to assign one new Message Type and one new TLV

6.1. New Message Type

New Message Type:

Value	Meaning
-----	-----
TBD	MPLS echo relay reply

6.2. New TLV

New TLV: Routable Relay Node Address TLV

Type	Meaning
----	-----
TBD	Relay Node Address Stack TLV

7. Acknowledgement

The authors would like to thank Carlos Pignataro, Xinwen Jiao and Manuel Paul for their valuable comments and discussions.

8. References

8.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC4377] Nadeau, T., Morrow, M., Swallow, G., Allan, D., and S. Matsushima, "Operations and Management (OAM) Requirements for Multi-Protocol Label Switched (MPLS) Networks", RFC 4377, February 2006.

- [RFC4378] Allan, D. and T. Nadeau, "A Framework for Multi-Protocol Label Switching (MPLS) Operations and Management (OAM)", RFC 4378, February 2006.
- [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.
- [RFC6425] Saxena, S., Swallow, G., Ali, Z., Farrel, A., Yasukawa, S., and T. Nadeau, "Detecting Data-Plane Failures in Point-to-Multipoint MPLS - Extensions to LSP Ping", RFC 6425, November 2011.

8.2. Informative References

- [ietf-mpls-seamless]
Leymann, N., Decraene, B., Filsfils, C., Konstantynowicz, M. and D. Steinberg, "Seamless MPLS Architecture", draft-ietf-mpls-seamless-mpls-00 , May 2011.

Authors' Addresses

Ryan Zheng (editor)
ZTE
50, Ruanjian Avenue
Nanjing, 210012, China

Email: zheng.zhi@zte.com.cn

Lizhong Jin (editor)
ZTE
889, Bibo Road
Shanghai, 201203, China

Email: lizhong.jin@zte.com.cn

Thomas Nadeau (editor)
Juniper

Email: tnadeau@juniper.net

George Swallow (editor)
Cisco
300 Beaver Brook Road
Boxborough , MASSACHUSETTS 01719, USA

Email: swallow@cisco.com

MPLS Working Group
Internet-Draft
Intended status: Standards Track
Expires: January 7, 2013

Katherine Zhao
Renwei Li
Huawei Technologies
Christian Jacquenet
France Telecom Orange
July 6, 2012

Fast Reroute Extensions to Receiver-Driven RSVP-TE for Multicast Tunnels
draft-zlj-mpls-mrsvp-te-frr-00.txt

Abstract

This document specifies fast reroute procedures to protect multicast LSP tunnels built by mRSVP-TE, a receiver-driven extension to RSVP-TE specified by [I-D.draft-lzj-mpls-receiver-driven-multicast-rsvp-te]. This document is motivated by the observation that the existing FRR solution specified by [RFC4090] and [RFC4875] for the sender-driven RSVP-TE is no longer applicable to the receiver-driven RSVP-TE.

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 7, 2013.

Copyright Notice

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

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must

include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

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

Table of Contents

1. Terminology	4
2. Introduction	4
2.1. Link Protection and Node Protection with mRSVP-TE	5
2.2. Primary and Backup LSP	8
2.3. Detour Backup and Facility Backup	8
3. Detour Backup for mRSVP-TE	9
3.1. Link Protection in Detour Backup Mode	9
3.1.1. Detour LSP Setup Scenario for Link Protection	9
3.1.2. Label Allocation for Link Protection	10
3.1.3. Link Failure Repair in Detour Mode	11
3.1.4. Re-convergence after Local Repair	12
3.2. Node Protection in Detour Backup Mode	12
3.2.1. Detour LSP Setup for Node Protection	12
3.2.2. Label Allocation and Binding for Node Protection	13
3.2.3. Node Failure Repair in Detour Mode	14
3.2.4. Re-Convergence after Local Repair	14
4. Facility Backup for mRSVP-TE	14
4.1. Link Protection in Facility Backup Mode	15
4.1.1. Backup LSP Setup for Link Protection	15
4.1.2. Label Allocation for Link Protection	15
4.1.3. Link Failure Repair in Facility Mode	17
4.1.4. Re-Convergence after Local Repair	18
4.2. Node Protection in Facility Backup Mode	18
4.2.1. Backup LSP setup in Facility Mode	18
4.2.2. Label Allocation for Node Protection	19
4.2.3. Node Failure Repair and Packet Encapsulation	22
4.2.4. Re-convergence after Local Repair	22
5. IANA Considerations	22
6. Manageability Considerations	22
7. Security Considerations	22
8. Acknowledgements	23
9. References	23
9.1. Normative References	23
9.2. Informative References	23
Authors' Addresses	24

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 RFC2119 [RFC-WORDS]. The reader is assumed to be familiar with the terminology in [RSVP], [RSVP-TE] and [mRSVP-TE].

This document uses same terminologies stated in [I-D.draft-lzj-mppls-receiver-driven-multicast-rsvp-te], [RFC4090], [RFC4875] and other FRR related IETF documents. In addition, some key notions and terminologies for this document are explained as follows:

- o mLSP, Multicast Label Switched Path, is either a P2MP or MP2MP LSP consisting of one or more sub-LSPs.
- o mRSVP-TE, Multicast Resource Reservation Protocol-Traffic Engineering, is used to distinguish from the regular sender-driven RSVP-TE. One major difference between RSVP-TE and mRSVP-TE is that the tunnel setup is initiated and driven by the data receiver instead of the data sender. The receiver-driven mRSVP-TE is best applicable to the setup of multicast LSP tunnels.
- o PLR: Point of Local Repair, an LSR that detects a local failure event and redirects traffic from protected mLSP to a backup mLSP tunnel which is supposed to locally repair the protected tunnel.
- o MP: Merge Point, an LSR that merges the traffic from backup tunnels with primary LSP at the forwarding engine. In the receiver-driven RSVP-TE for multicast tunnels, MP is the LSR that initiates backup mLSP setup taking PLR as the root of backup LSR.
- o N: The node to be protected.
- o Pn: The node(s) on the backup path for protecting the node N.
- o Root: A router where an mLSP is rooted at. Data enters the root and then is distributed to leaves along the P2MP/MP2MP LSP.
- o FRR Domain: A set of links and LSRs cross over a protected sub-LSP and backup LSP, which is between PLR and MP(s).

2. Introduction

Fast Reroute technology has been well accepted and deployed to provide millisecond-level protection in case of node/link failures.

FRR employs some local repair mechanisms to meet the fast reroute requirements by computing and provisioning backup tunnels in advance of failure and by redirecting traffic to such backup tunnels as close to the failure point as possible.

The fast reroute extensions to RSVP-TE are specified in [RFC4090] and [RFC4875]. Such extensions work well with the sender-driven RSVP-TE, but they are no longer applicable to the receiver-driven RSVP-TE for multicast tunnels described in the draft [I-D.draft-lzj-mppls-receiver-driven-multicast-rsvp-te].

In the receiver-driven paradigm of mRSVP-TE, the procedure to set up an LSP tunnel is inverted from that in the sender-driven RSVP-TE, and thus the backup mLSP setup and failover handling mechanism will have to be different from what has been specified for the sender-driven RSVP-TE. From the signaling point of view, the behavior of PLR and MR are inverted from the sender-driven paradigm of RSVP-TE, the setup for a backup mLSP is initiated by MP with PLR being taken as the root of a P2MP/MP2MP tree. The RSVP PATH message is sent from MP towards PLR with the FAST_REROUT, DETOUR as well as other FRR related objects conveyed in the PATH message. RSVP RESV message is sent from PLR towards MP carrying FRR information such as the inner label used to represent a protected mLSP tunnel, etc.

On the other hand, from the packet forwarding point of view, the behavior of PLR and MP are not inverted comparing to the sender-driven RSVP-TE. The traffic switchover and redirecting are still initiated by PLR, and the data traffic is merged at MP in the same way as what is specified for the sender-driven RSVP-TE.

This document will describe various FRR protection methods and behavior changes for the receiver-driven mRSVP-TE, and specify fast-reroute extensions to the RSVP-TE messages, mechanisms and procedures specified in the mRSVP-TE draft [I-D.draft-lzj-mppls-receiver-driven-multicast-rsvp-te].

2.1. Link Protection and Node Protection with mRSVP-TE

FRR link protection aims to protect a direct link between two LSRs (Label Switch Routers). An LSR at one end of the link is called PLR (Point of Local Repair), and the other LSR on the other end of the link is called MP (Merge Point). A backup LSP whose setup is originated at MP and terminated at PLR will be established to protect the primary LSP crossing over the link. The LSR over the backup path is called Pn. These connected LSRs and links are called an FRR domain in this document. An example of an FRR domain supporting link protection is shown in Figure 1.

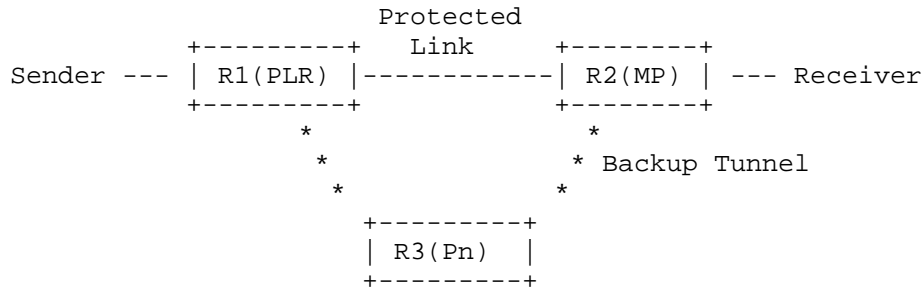


Figure 1: Basic FRR Link Protection

In an FRR domain constructed by mRSVP-TE, MP initiates both the primary and the backup LSP setup at the signaling control plane, and merges the traffic from the backup LSP into the primary LSP at the data forwarding plane. The PLR works with the MP to set up LSP at the signaling control plane accordingly, and detects link failure and initiates local repair at the data forwarding plane. In Figure 1, we use hyphen - to denote a primary tunnel between LSRs; use asteroid * to denote a backup tunnel. The same symbols will be applied to all figures throughout the document.

Node protection is a technique used to protect a node N that resides between PLR and MP over a primary LSP. An example of node protection is shown at Figure 2.

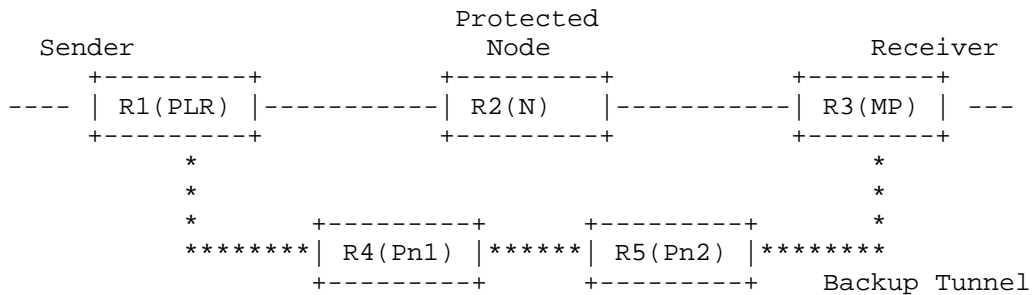


Figure 2: Basic FRR Node Protection

N (R2) denotes a node being protected over a primary LSP, its upstream node plays the role of PLR and downstream node plays the role of MP. Pn denotes a transit node over its backup LSP. Note that there can be multiple Pn's over a backup tunnel. Pn does not play a significant role for FRR but works as a regular LSR to receive and transmit multicast data and signaling messages over backup LSPs.

Besides the basic P2P node protection, mRSVP-TE is more interested on the P2MP and MP2MP node protection, as shown at Figure 3 and Figure 4. Because the same protection mechanism can be commonly used for both P2MP and MP2MP, this document uses P2MP as example for the discussion, and mention MP2MP only if there is a difference from P2MP.

There are two typical methods to protect a P2MP multicast tree, one uses a P2MP tree as a backup LSP to protect a primary mLSP (see Figure 3), and the other uses multiple P2P LSPs to protect a P2MP mLSP (see Figure 4).

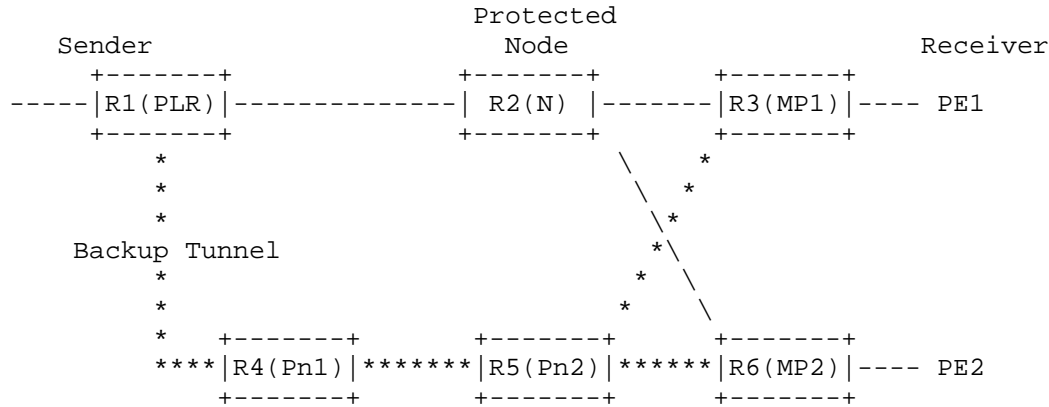


Figure 3: P2MP Node Protection in Facility Mode

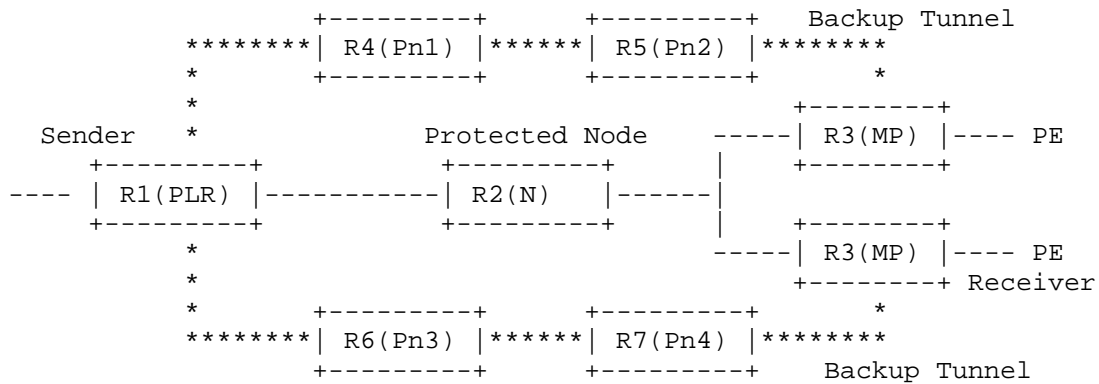


Figure 4: Multiple P2Ps Protecting a P2MP LSP

2.2. Primary and Backup LSP

A router that experiences a node/link failure must have pre-determined which alternate reroute path to protect such a failure. The alternate backup path should be established before a protected LSP is broken. Anything such as backup route computation and configuration required for local repair should be done prior to failure occurrence so that the failover time can be reduced to minimum.

On the control plane, the backup LSP will be set up along with its primary LSP setup. The PATH/RESV refresh messages are transmitted over both protected and backup LSPs before failover. However on the data plane, there are two implementation options for traffic forwarding. One option is that the user traffic does not transmit on backup LSP tunnel until a failure is detected and the local repair takes place. The second option is to have user traffic transmitted on both protected and backup mLSPs before failover, LSR at Merge Point will drop the packets from backup path before switchover. The second option can further reduce traffic switchover time but causes more overhead. This document leaves the flexibility for implementation to decide which option to choose, but will use the first option for the discussion, i.e. we assume that the traffic only transmits on the primary LSP before switchover.

2.3. Detour Backup and Facility Backup

Due to historic reasons and implementation preferences, two independent methods of doing fast reroute have been developed. One backup method is called detour backup that is specially designed for 1:1 protection. And the other one is called facility backup that is specially designed for 1: N protection, where N can be equal to or more than 1. From the point of view of applications, the facility backup method can support both 1:N and 1:1, but from the technical point of view, they are two different methods requiring different implementations with respect to their label stacks when forwarding packets.

The detour backup creates a dedicated LSP to protect an LSP and uses a single MPLS label for packet encapsulation; its implementation is simpler but consumes more label resources. The facility backup creates a common LSP to protect a set of LSPs that have similar backup constraints, this method takes advantage of MPLS label stacking and uses dual-label encapsulation, thus it can save some label resources compared to the detour backup method.

These two solutions have co-existed as options for vendors and service providers to choose. This document will specify both the

methods. Throughout the document, the detour method is used to represent 1:1 protection while facility method is used to represent 1: n protection. The term detour LSP is specially used for 1:1 protection while backup LSP is used for 1: N protection, but sometimes the later one can be used for common cases when no ambiguity arises.

3. Detour Backup for mRSVP-TE

This section specifies mechanisms and procedures for mRSVP-TE fast reroute by using the detour backup method. The term detour LSP will be used to represent the LSP in the detour mode and for one-to-one protection without special remark.

3.1. Link Protection in Detour Backup Mode

3.1.1. Detour LSP Setup Scenario for Link Protection

A detour LSP setup is initiated by MP along with the setup of the protected LSP (refer to Figure 1 for the topology), which is one of the major differences from the procedure stated in [RFC4090] and [RFC4875]. Following the LSP setup procedure specified by the draft [I-D.draft-lzj-mpls-receiver-driven-multicast-rsvp-te], MP sends RSVP PATH messages towards the sender over a primary path. For the link protection, MP and PLR are directly connected by the link being protected, hence the PATH message is sent from MP to PLR directly upstream.

MP is not necessarily the originator of the primary LSP, but is the first LSR entering an FRR domain along the primary route, and thus our discussion for LSP setup starts from MP.

Once the PATH message is sent out, MP will check to see if there is detour route available for the detour link protection. The detour route calculation can be done by running CSPF on the link state database produced by IGP protocols with TE extensions. There is no change required for backup route computation, and the LSP computation will be based on this assumption without additional explanation.

If the CSPF stack returns 'no detour route found' after the calculation, MP stops the detour LSP setup and traverses to the NHOP over the primary path. It considers NHOP as another MP and starts the FRR process again. If at least one detour route is found by CSPF, MP selects the shortest route and initiates the detour LSP setup. MP considers PLR as the end point of detour LSP and sends a PATH message towards PLR hop by hop. In the example of Figure 1, the PATH message will be sent to Pn (R3) and then relayed to PLR (R1).

PLR replies such a PATH message with a RESV message towards MP through Pn(s). The transit node Pn(s) just relay the PATH/RESV messages without any special process required for the link protection. Detour LSP setup is done once RESV is received and processed by MP.

3.1.2. Label Allocation for Link Protection

Because the detour method uses a dedicated backup LSP to protect a primary LSP, one-to-one binding can be made for a pair of primary and backup LSPs, a single MPLS label encapsulation will be sufficient for packet forwarding and local failure repair. DLA (downstream label allocation) can be used as the label assignment method over the detour tunnel for the link protection. With mRSVP-TE, a downstream label is assigned by an LSR that is sending PATH message to its upstream router, and an upstream label is assigned by an LSR that is sending the RESV message to its downstream router. The label allocation, however, is more complicated when the primary LSP is a P2MP or MP2MP tree. A special upstream label allocation and resource preemption method is introduced and discussed to handle the protection for P2MP and MP2MP tree structures in a later section.

An example of the label allocation for link protection in the detour mode is given in Figure 5. For the sake of readability, we use label Lp to represent the label assigned to the primary tunnel, label Lb assigned to the backup tunnel. For example, Lp2 represent a downstream label assigned for LSR R2 to receive incoming data over the primary tunnel. Lb2 represents a downstream label assigned for R2 to receive data over a detour LSP.

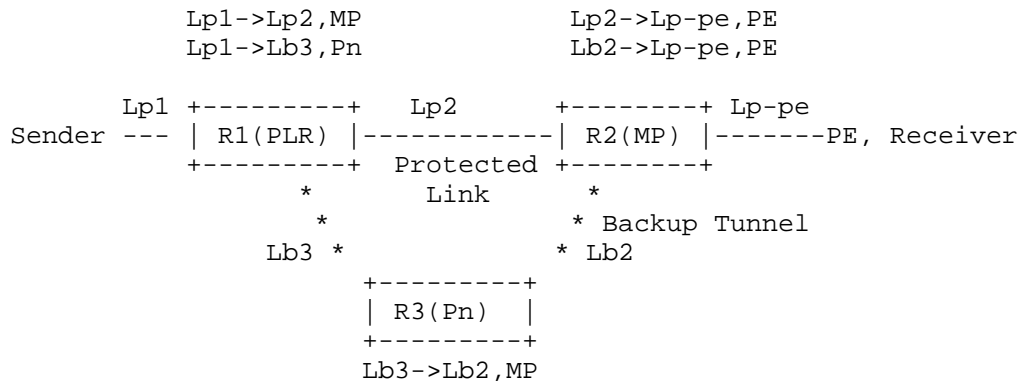


Figure 5: Label Allocation for Link Protection in Detour Mode

In the example of Figure 5, MP assigns label Lp2 and sends it to PLR

via the PATH message over the link {MP-PLR} to set up the primary LSP. For the detour route {MP-Pn-PLR}, MP assigns a label Lb2 and sends it to Pn via the PATH message. MP binds label Lp2 with label Lb2 for this pair of the primary and detour LSPs. An entry 'Lp2->Lp-pe, PE' will be added into MP's FIB for packet forwarding over the protected LSP. Another entry 'Lb2-> Lp-pe, PE' will be added and used when traffic is received from the detour tunnel upon switchover.

Pn (transit node) on the detour tunnel receives Lb2 from MP. Pn assigns a downstream label Lb3 and sends it to PLR via a PATH message. Pn will add an entry 'Lb3->Lb2, MP' to its FIB for packet forwarding. Note that Pn is not aware of the primary traffic so there is only one forwarding entry needed in its FIB.

PLR receives two PATH messages from MP and Pn respectively. Then it binds label Lp2 from the primary LSP with label Lb3 from the detour LSP. The detour LSP ends at PLR while the primary LSP may not end at PLR if PLR is not the root of the P2MP tree. PLR will allocate a downstream label Lp1 and sends it to its upstream router, which is outside of the FRR domain in this example thus not shown at Figure 5. There will be two entries added into PLR's FIB: one entry 'Lp1->Lp2, MP' for the primary traffic forwarding, and the another entry 'Lp1->Lb3, Pn' for the detour traffic forwarding upon failover.

PLR processes PATH messages from MP and sends RESV messages towards MP. If the primary sub-LSP is over a P2MP tree, PLR will not allocate upstream labels for receiving traffic from the downstream node (MP or Pn in this example) because the traffic is unidirectional. If the sub-LSP is over an MP2MP tree, PLR will allocate an upstream label for receiving traffic from opposite direction, Pn(s) will repeats the same and allocate upstream label for MP2MP. Detour LSP setup is completed once MP has received and processed the RESV message originated by PLR. Figures 5 shows the summary of labels allocated and FIB entries created on each node in the FRR domain.

3.1.3. Link Failure Repair in Detour Mode

Link failure can be detected by, for example, BFD (Bidirectional Forwarding Detection) along the protected LSP. The failure detection algorithm is the same as what is used for the sender-driven RSVP-TE.

Once a link failure is detected by PLR and all switchover criteria are met, PLR will redirect the traffic to the detour LSP based on the forwarding entry 'Lp1->Lb3, Pn'. The entry 'Lp1->Lp2, MP' for primary path will be deactivated.

Pn works as a normal label switch router and forward MPLS packet to MP. MP receives the packet and figures out that the packet is from the detour path, so the packet will be forwarded to PE based on the entry 'Lb2->Lp-pe, PE'. The detour traffic is therefore merged back to the primary LSP towards PE, which completes the link failure repairing by detouring and merging the traffic.

3.1.4. Re-convergence after Local Repair

Routers outside of the FRR domain are not impacted by the link failure and local repair for the protection mechanism discussed in the previous sub-sections. Traffic is transmitted over a detour LSP after a link failure and local repair. Usually the detour path is not the shortest path so the network will eventually re-converge and a new shortest path will be calculated by the MPLS control plane. Once a new primary path is determined, the traffic is no longer transmitted through the detour LSP and PLR will be notified to tear down the detour LSP and clean up its internal stack. PLR will send a PathTear message to Pn and MP for tearing down the detour LSP and release backup labels. Re-convergence procedure is the same as the procedure used for sender-driven RSVP-TE FRR.

3.2. Node Protection in Detour Backup Mode

3.2.1. Detour LSP Setup for Node Protection

The detour LSP setup for the node protection is similar to the link protection. Take Figure 2 as an example, where the node N being protected resides between MP and PLR. In this case the two sub-links {MP-N} and {N-PLR} are also to be protected in conjunction with the node N protection. It is assumed that the link protection mechanism given in the previous sub-section is applicable to the sub-link protection in this situation. Hence this section will focus on the procedure to handle the node protection. A combined solution for providing the node protection in conjunction with the link protection can be derived from the discussions in section 3.1 and this section.

For the node protection shown in Figure 2, MP(R3) sends a PATH message to N for the primary LSP setup, the primary LSP in the FRR domain goes through the route {MP-N-PLR}. Once the PATH message is sent out to N, MP checks to see if there is a detour path available for node N by using CSPF computation, which would indicate N as a node to be avoided on the detour path. If no detour route is found, skip the detour LSP setup. If a detour route is found, MP initiates the detour LSP setup and consider PLR as the terminator of the detour LSP. MP sends a PATH message towards PLR over the detour route hop by hop, in the example of Figure 2, the detour route is in the order of {MP-Pn2-Pn1-PLR}. Similar to the link protection, PLR sends back

RESV message towards MP through Pn(s). Transit node Pn(s) just relay the PATH and RESV messages without special processes required for the node protection. The detour LSP setup is completed once the RESV message is received and processed by MP.

Figure 2 shows a case of the basic node protection where N is not a branch node; it will be more complicated when N is a branch node over a P2MP / MP2MP tree structure. The mechanism for these cases will be given later in section 5.2.2.

3.2.2. Label Allocation and Binding for Node Protection

Same as the link protection, the node protection uses the single label encapsulation and downstream label allocation method in the detour backup mode. An example of the label allocation for the node protection is given in Figure 6.

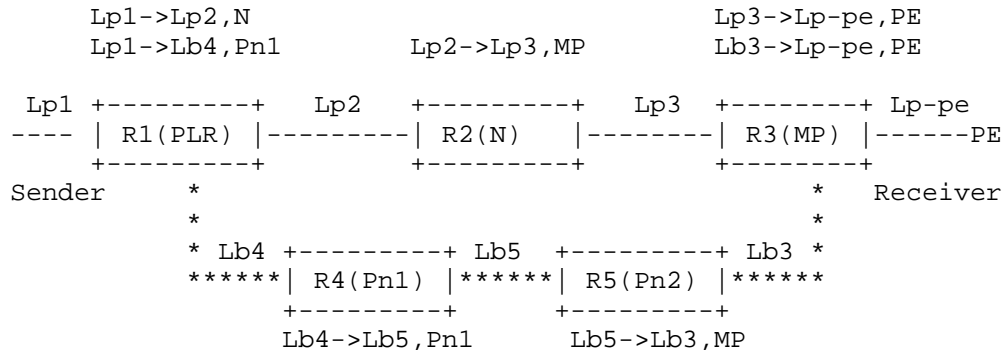


Figure 6: Node Protection in Detour Mode

MP (R3) assigns a label Lp3 for the primary LSP and sends it to node N via a PATH message over the protected route {MP-N-PLR}, N will allocate a downstream label Lp2 and sends it to PLR via a PATH message. MP also assigns a label Lb3 for the detour LSP and sends it to Pn2 via a PATH message over the detour route {MP-Pn2-Pn1-PLR}. MP binds label Lp3 with label Lb3 for this pair of primary and backup LSP. An entry 'Lp3->Lp-pe, PE' will be added to MP's FIB for primary packet forwarding over the protected LSP. Another entry 'Lb3->Lp-pe, PE' will be kept in the FIB and used when a failover takes place and traffic is redirected to the detour LSP.

There could be multiple transit nodes Pn(s) along the detour LSP, each of which will allocate a downstream label and sends it to its upstream router. Eventually PLR receives the PATH message from the

protected node N and the transit node Pn1 in this example. PLR binds primary label Lp2 with the detour label Lb4, and adds two entries into its FIB: One entry 'Lp1->Lp3, N' for the primary traffic forwarding, and another entry 'Lp1->Lb4, Pn1' for the detour traffic forwarding. The allocated labels and FIB entries in the FRR domain can be found in Figure 6.

3.2.3. Node Failure Repair in Detour Mode

Once the node N failure is detected by PLR, it will redirect the traffic from the primary LSP to its detour LSP based on the binding and forwarding entry 'Lp1->Lb4, Pn1'. The data packet is forwarded through LSR->Pn1-Pn2->MP. Eventually MP will receive the packet from the detour path. Consulting its FIB forwarding entry 'Lb3->Lp-pe, PE', the data packet will be forwarded to PE, therefore the detoured traffic gets merged into the primary path.

The local repair mechanism for the node protection is the same as the link protection in the detour mode except that there are two links {MP-N} and {N-PLR} to be protected in conjunction with the node N protection. The FRR domain must be configured so that both the link failure detection and node failure detection methods are specified. For example, BFD may be used for this purpose and are configured as follows:

- o BFD1 between MP and N;
- o BFD2 between N and PLR;
- o BFD3 between PLR and MP;

PLR and MP can apply either link repair or node repair or both depending on the results of BFD detection.

3.2.4. Re-Convergence after Local Repair

After a node failure takes place, the network topology will change. And therefore the network will eventually re-converge and a new best path will be found the primary LSP. PLR will be notified as soon as the new primary path is signaled and set up. PLR will send notification message to Pn1 and MP for tearing down the detour LSP and withdraw backup labels.

4. Facility Backup for mRSVP-TE

This section specifies mechanisms and procedures for mRSVP-TE fast reroute by using the facility backup method. The term backup LSP

will be used to represent the LSP in the facility mode for 1: N protection without any special remark. Note that the term 'detour LSP' is no longer used in this section for the Facility backup.

The backup LSP differs from the detour LSP in that one single backup LSP is used to protect multiple primary LSPs. General speaking, two labels will be used for the backup LSP with the inner label being used to indicate which primary LSP is being protected.

4.1. Link Protection in Facility Backup Mode

4.1.1. Backup LSP Setup for Link Protection

Same as in the detour LSP setup, MP sends a RSVP PATH message towards PLR over the primary route. Once the PATH message is sent out, MP will execute the backup LSP procedures in the following steps:

- o Check if there has been a backup LSP created to protect the link between PLR and MP. If a backup LSP is found, skip the further process at MP, e.g. does not send a PATH message over the backup route for LSP setup. However it does not mean that no process is needed for the link protection. Later on PLR will allocate an inner label for each newly created primary LSP and send it to Pn(s) and MP via the RESV message. The details for label allocation and packet encapsulation will be discussed in the next section 4.1.2.
- o If there is no existing backup LSP available, MP initiates the backup LSP setup: MP calculates a backup route by using CSPF by taking PLR as the endpoint of the back LSP and sends a PATH message towards PLR hop by hop over the backup route. In the example of Figure 1, PATH will be sent from MP to Pn (R3) and relayed to PLR (R1). PLR will then send MP a RESV message to complete the backup LSP setup. The next sub-section will specify the details about the label allocation and binding.

4.1.2. Label Allocation for Link Protection

As a backup LSP protects one or more primary LSPs, Facility Protection uses dual-label for packet forwarding, in which the outer label is used for regular packet forwarding hop by hop over the backup LSP while the inner label is used to represent a primary LSP and used by MP to merge the backup traffic to its corresponding primary LSP. Multiple primary LSPs will share the common outer label while the inner label is unique for each protected LSP. Figure 7 below shows how dual-label stack is assigned and used for the facility backup. There are two primary LSPs to be protected by a common backup LSP in this example.

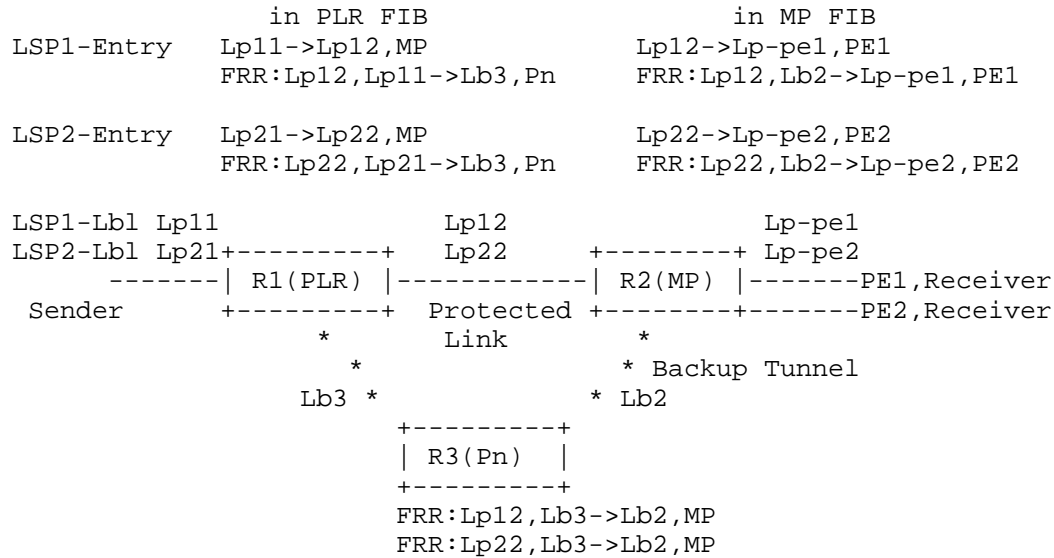


Figure 7: Label Allocation for Link Protection in Facility Mode

Assume that the primary LSP1 is created first, MP assigns a downstream label Lp12 for LSP1 being protected and sends the label to PLR via a PATH message over route {MP-PLR}. Because the primary LSP1 is the first LSP created over this route, MP also assigns a downstream label Lb2 for the backup LSP and sends it to Pn via a PATH message over the backup route {MP-Pn-PLR}. Pn allocates a downstream label Lb3 and sends it to PLR via a PATH message.

Once PATH messages are received from MP and Pn respectively, PLR will allocate an inner label to represent the primary LSP1 for the backup LSP. The method to allocate the inner label is up to implementation. In this example, label Lp12 is used as the inner label to represent primary LSP1 over the backup LSP. LSR at merge point uses the inner label to locate the corresponding primary LSP. The inner label is propagated from PLR to MP by a RESV message. Note that PLR and MP are the LSRs that actually see, use or process the inner label, while other transit node Pns do not process the inner label.

The process for the second or more primary LSPs protected by the same backup LSP is different from that for the first one. MP does not allocate any new downstream label for the backup LSP since the backup LSP for the first primary LSP is shared between all the primary LSPs protected by the same backup LSP. But the PLR is required to allocate an inner label for each newly created primary LSP and sends it to MP hop by hop via a RESV message.

We use Figure 7 as an example to show the packet forwarding FIB entry by using the following format:

```
FRR:(inner label),(incoming outer label)->(outgoing outer label),NHOP
```

When MP allocates the downstream label Lp12 for the primary LSP1, an entry 'Lp12->Lp-pe1, PE1' is added into MP's FIB. Another FRR entry 'FRR: Lg12, Lb2->Lp-pe1, PE1' is added when MP receives a RESV message that carries an inner label Lg12 and binding information with LSP1. So the MP will have two forwarding entries for each protected LSP. In this example MP will have four entries in FIB for the two protected paths LSP1 and LSP2:

```
Lp12->Lp-pe1, PE1
```

```
Lp22->Lp-pe2, PE2
```

```
FRR: Lp12, Lb2 -> Lp-pe1, PE1
```

```
FRR: Lp22, Lb2 -> Lp-pe2, PE2
```

PLR creates a forwarding entry for a primary LSP whenever it receives a PATH message for the setup of a new primary LSP. For each primary path LSP1, once PLR receives the PATH message from the backup route, PLR allocates an inner label for the primary LSP and creates an FRR entry in FIB. PLR FIB will have these entries for the two protected LSP LSP1 and LSP2:

```
Lp11 ->Lp12, MP
```

```
Lp21->Lp22, MP
```

```
FRR: Lp12, Lp11 -> Lb3, MP
```

```
FRR: Lp22, Lp21 -> Lb3, MP
```

Note that the transit routers Pn uses the outer label for packet forwarding and keeps the inner label untouched.

4.1.3. Link Failure Repair in Facility Mode

Before a link failure is detected, PLR encapsulates user packets with a single label Lp1 and forwards the packet to MP. MP also uses a single label encapsulation and forwards the packet to PE.

After a link failure is detected, PLR, for example, R1 in Figure 7, will encapsulate user packets with dual-label stack with outer label Lb2 used for packet forwarding in the backup path and inner label Lp2

used to map to the corresponding primary LSP. MP will pop out outer label Lb2 if needed, swap inner label Lp12 with Lp-pe1, and then forward the packet to PE1.

4.1.4. Re-Convergence after Local Repair

After a link failure occurs, the network will reconverge. PLR will be notified as soon as a new best path for the primary LSP will be found and activated. Then PLR will tear down the backup LSP, release backup labels and clean up entries in the FIB.

4.2. Node Protection in Facility Backup Mode

4.2.1. Backup LSP setup in Facility Mode

Two methods for node protection in facility mode have been illustrated in Figure 3 and Figure 4. The method shown in Fig 3 uses a P2MP or MP2MP backup LSP to protect a branch node N; the method shown in Fig 4 uses multiple LSPs to protect the node N. It is seen that the first method can reduce the traffic replication on the backup LSP; the second method suffers from traffic overhead because multiple backup sub-LSPs are used. Which method to use is an implementation option. In this document we will use the method shown in Fig 3 to describe the node protection mechanism in the facility mode.

Some special processes are needed for the P2MP or MP2MP tree setup and label allocation. Assume that LSR PE1 joins a primary P2MP tree structure in the example of Fig 3. PE1 sends a RSVP PATH message to MP1 for LSP setting up, this PATH message will be relayed to PLR through node N being protected. MP1 calculates the backup route with a constraint to avoid the node N; it initiates the backup LSP setup by sending a PATH message over the backup path {MP1-Pn2-Pn1-PLR}. RSVP RESV messages will be replied by PLR to MP1 through the primary route {PLR-N-MP1} and the backup route {PLR-Pn1-Pn2-MP1} respectively.

Later on, another LSR PE2 joins the P2MP tree by sending a PATH message to MP2. MP2 will relay the PATH message to the node N being protected. Now N becomes a branch node and the PATH message sending to PLR can be suppressed. MP2 performs the same procedure as MP1 did for the first branch {PE1-MP1-N}, a backup route {MP2-Pn2-Pn1-PLR} will be found by CSPF calculation, the node Pn2 now becomes a branch node crossing over the backup P2MP tree. The PATH message suppose to send from Pn2 to PLR can be suppressed by the branch node Pn2. RSVP RESV messages will be replied by PLR to MP2 through the primary route {PLR-N-MP2} and the backup route {PLR-Pn1-Pn2-MP2} respectively.

Whenever the second or more primary LSP(s) are added going through the same node N and PLR, all these primary LSPs can be protected by the single backup LSP. The procedure to setup the primary LSP is the same as what is used for the first primary LSP setup, the key technique is to allocate unique identifier of a primary LSP and bind it with the backup LSP, the mechanism will be discussed in the next sub-section.

4.2.2. Label Allocation for Node Protection

In order to achieve 1:n protection in Facility mode, a unique identifier must be assigned to represent each primary LSP being protected. This identifier should be advertised to all LSRs in a FRR domain and used for traffic switchover in case of node N failure. There are many ways to assign and use the identifier, this document gives an sample mechanism about how to use ULA (upstream label allocation) to assign a MPLS label and apply it as the identifier of a primary LSP. Figure 8 gives an example of label allocation and FIB entry creation for the node protection in Facility mode.

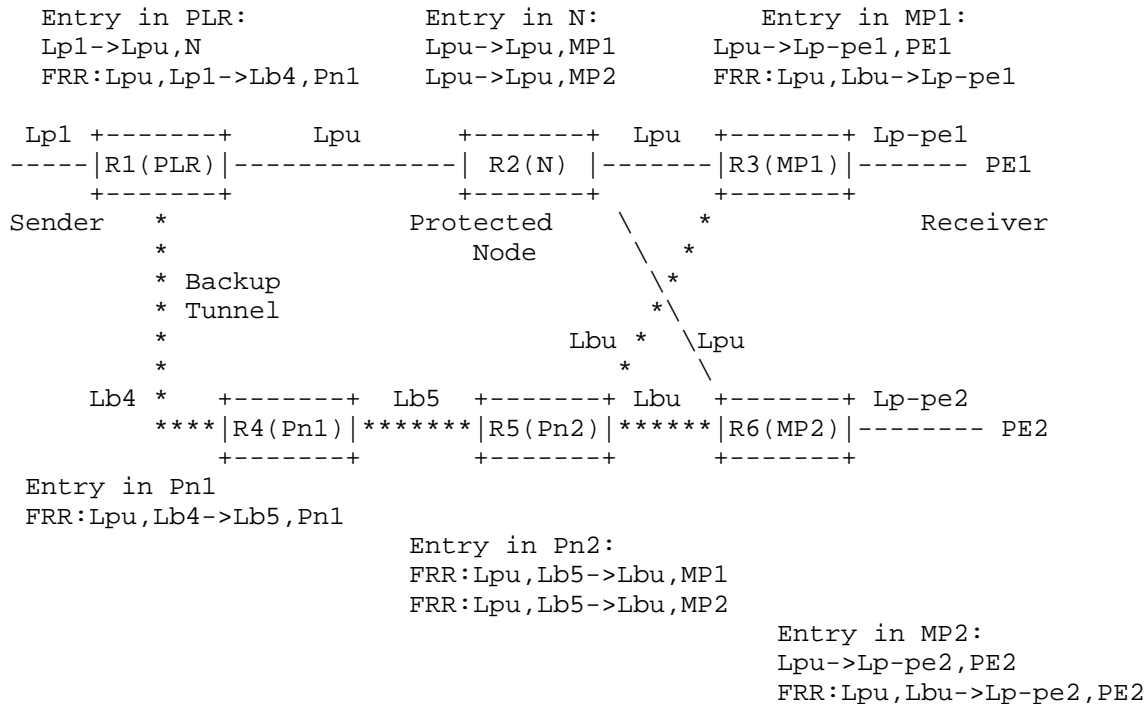


Figure 8: Label Allocation for P2MP Node Protection in Facility Mode

In the FRR domain of Figure 8, an identical label Lpu is assigned to

these sub-LSPs over the primary LSP: {PLR-N}, {N-MP1} and {N-MP2}. Lpu can be allocated by the branch node N for the primary LSP and used as the identifier of the primary LSP. If there are multiple primary LSPs crossover the same node N and to be protected by the single backup LSP, there will be multiple Lpu labels assigned for each of the primary LSP accordingly. In order to guarantee the uniqueness of Lpu in node N and MPs, the LSRs are required to have ULA capability in FRR domain. In addition an algorithm for ULA assignment and negotiation among the LSRs will need to be further specified by the later IETF draft.

During normal operation, PLR encapsulates sender's packet with the label Lpu and forwards the packet to the node N over the primary LSP. The node N as a branch node will replicate the traffic to MP1 and MP2 using label Lpu in this example. When a node failure is detected PLR will redirect the traffic to the backup LSP, and the dual-label stack will be used for packet encapsulation over the backup LSP, where the inner label is Lpu to represent a primary LSP; the outer label is allocated by MP and Pn(s) using DLA (downstream label allocation), which is used for packet forwarding over backup LSP via regular RSVP-TE mechanism.

Detailed label allocation on each LSR is described at below.

1. Label Allocation and FRR Entry on MP1 and MP2:

For the first primary LSP setup, MP1 assigns a downstream label Lpdla for the primary LSP and sends it to the protected node N via PATH message. The node N discards Lpdla and uses ULA to assign a new label Lpu that will be used as a downstream label for N to send packet to MP1.

Node N sends the label Lpu to MP1 via RESV message; MP1 replaces its downstream assigned label Lpdla with Lpu. If Lpu has been used by other tunnel on the LSR, MP1 will request the node N to reassign the Lpu. In case of conflict an ULA negotiation procedure has to be executed (this procedure is TBD).

MP1 too assigns a downstream label Lbdla for the backup LSP and sends it to Pn2 via PATH message over the backup route {MP1-Pn2-Pn1-PLR}. Pn2 is a branch node so it will perform the same procedure as the branch node N on the primary LSP. Pn2 discards the label Lbdla received from the PATH message, assigns a new label Lbu and sends it to MP1 via RESV message.

Once a RESV message is originated by PLR and sent through the backup route, MP1 will get an inner label Lpu that represents the primary LSP in this example. MP1 adds FRR entry with both inner and outer

label. MP1 FIB will have two forwarding entries for the LSP being protected in Facility mode:

Lpu->Lp-pe1, PE1

FRR: Lpu, Lbu->Lp-pe2, PE2

With the same process, MP2 will have two forwarding entries for the LSP being protected:

Lpu->Lp-pe2, PE2

FRR: Lpu, Lbu->Lp-pe2, PE2

2. Label Allocation and FRR Entry on Pn2 and Pn1:

As mentioned in the last paragraph, when Pn2 (transit branch node) receives PATH message from MP1 and MP2 respectively, it will allocate label Lbu and sends to each MP. Pn2 will have two forwarding entries for the LSP being protected:

FRR: Lpu, Lb5->Lbu, MP1

FRR: Lpu, Lb5->Lbu, MP2

Pn1 is a transit node and has only one FRR entry for the LSP being protected:

FRR: Lpu, Lb4->Lb5, Pn2

3. Label Allocation and FRR Entry on PLR:

PLR receives a PATH message from the node N that carries a downstream label Lpu; and a PATH message from Pn1 that carries a downstream label Lb5. PLR uses Lpu as an inner label for the primary LSP and sends it to Pn1 towards MPs via RESV message. PLR will have two entries for a LSP being protected:

Lp1->Lpu, N

FRR: Lpu, Lp1->Lb1, Pn1

For every add-in primary LSP being protected by the same backup LSP, PLR will assign an inner label and send it to LSRs cross the backup LSP so that each of LSR can add corresponding FRR entry into FIB and use them for traffic switchover during local repair.

4.2.3. Node Failure Repair and Packet Encapsulation

Once protected node N fails and the failure is detected by PLR, it will initiate a switchover by redirecting the traffic to backup LSP. Packet encapsulation in each LSR over the backup LSP will be done based on the FRR entries in its FIB. For example a packet arrived on PLR suppose to be forwarded to node N by using entry 'Lp1->Lpu, N', now should be forwarded to Pn1 based on entry 'FRR: Lpu,Lp1->Lb4, Pn1'. PLR encapsulates the packet with Lpu as inner label, Lb4 as outer label and forwards it to Pn1. Pn1 will swap outer label for packet forwarding and keep inner label untouched.

Once the packet reaches MP1, it will pop out the outer label, swap inner label with outgoing label Lp-pel and forward the packet to NHOP PE1 with a single label Lp-pel, the packet de-capsulation / encapsulation is based on the entry 'FRR: Lpu, Lbu->Lp-pel, PE1'. The dual label stack packet is terminated at MP1 and the traffic is merged with the primary path. The same procedure is applicable to receiver LSR MP2.

4.2.4. Re-convergence after Local Repair

Routers outside of FRR domain are not impacted by the link failure and local repair. However the network will eventually re-converge and a new best path to the sender or root will be found by PE1 and PE2. PLR will be notified as soon as the new primary path is determined. PLR will send notification message to Pn and MP for tearing down the detour LSP and withdraw backup labels. There is no difference between facility and detour method in terms of re-convergence process.

5. IANA Considerations

TBD.

6. Manageability Considerations

TBD.

7. Security Considerations

TBD.

8. Acknowledgements

We would like to thank Quintin Zhao, Lin Han, Emily Chen, and Robert Tao for discussions and comments.

9. References

9.1. Normative References

- [I-D.lzj-mppls-receiver-driven-multicast-rsvp-te]
Li, R., Zhao, Q., and C. Jacquenet, "Receiver-Driven Multicast Traffic Engineered Label Switched Paths", draft-lzj-mppls-receiver-driven-multicast-rsvp-te-00 (work in progress), March 2012.
- [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.
- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC3031] Rosen, E., Viswanathan, A., and R. Callon, "Multiprotocol Label Switching Architecture", RFC 3031, January 2001.
- [RFC3209] Awduche, D., Berger, L., Gan, D., Li, T., Srinivasan, V., and G. Swallow, "RSVP-TE: Extensions to RSVP for LSP Tunnels", RFC 3209, December 2001.
- [RFC2205] Braden, B., Zhang, L., Berson, S., Herzog, S., and S. Jamin, "Resource ReSerVation Protocol (RSVP) -- Version 1 Functional Specification", RFC 2205, September 1997.

9.2. Informative References

- [RFC3468] Andersson, L. and G. Swallow, "The Multiprotocol Label Switching (MPLS) Working Group decision on MPLS signaling protocols", RFC 3468, February 2003.
- [RFC3473] Berger, L., "Generalized Multi-Protocol Label Switching (GMPLS) Signaling Resource Reservation Protocol-Traffic Engineering (RSVP-TE) Extensions", RFC 3473, January 2003.

[RFC3564] Le Faucheur, F. and W. Lai, "Requirements for Support of Differentiated Services-aware MPLS Traffic Engineering", RFC 3564, July 2003.

Authors' Addresses

Katherine Zhao
Huawei Technologies
2330 Central Expressway
Santa Clara, CA 95050
USA

Email: katherine.zhao@huawei.com

Renwei Li
Huawei Technologies
2330 Central Expressway
Santa Clara, CA 95050
USA

Email: renwei.li@huawei.com

Christian Jacquenet
France Telecom Orange
4 rue du Clos Courtel
35512 Cession Sevigne,
France

Email: christian.jacquenet@orange.com