

Internet Engineering Task Force
Internet-Draft
Intended status: Standards Track
Expires: August 29, 2013

H. Chen
Huawei Technologies
N. So
Tata Communications
A. Liu
Ericsson
L. Liu
UC Davis
February 25, 2013

Extensions to RSVP-TE for P2MP LSP Ingress Local Protection
draft-chen-mppls-p2mp-ingress-protection-08.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 August 29, 2013.

Copyright Notice

Copyright (c) 2013 IETF Trust and the persons identified as the document authors. All rights reserved.

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

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

Table of Contents

1. Introduction	3
2. Terminology	3
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. Protocol Extensions	7
6.1. New RSVP-TE Messages	8
6.1.1. LSP Information Message	8
6.1.2. Backup LSP for One-to-One Backup	9
6.1.3. Backup LSP for Facility Backup	10
6.1.4. LSP Information Confirmation Message	11
6.2. New RSVP-TE Objects	12
6.2.1. Information about Existing LSP	12
6.2.2. Desire for Locally Protecting Ingress	12
6.2.3. Backup LSP for One-to-One Backup	13
6.2.4. Backup LSP for Facility Backup	13
6.3. OSPF Opaque LSA	14
6.4. Mapping Traffic to Backup LSP	14
7. IANA Considerations	15
8. Acknowledgement	15
9. References	15
9.1. Normative References	15
9.2. Informative References	16
Authors' Addresses	16

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. 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 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 sub tree, 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 (the primary ingress of the LSP) and Ra (the backup ingress node designated to protect the primary ingress). 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.

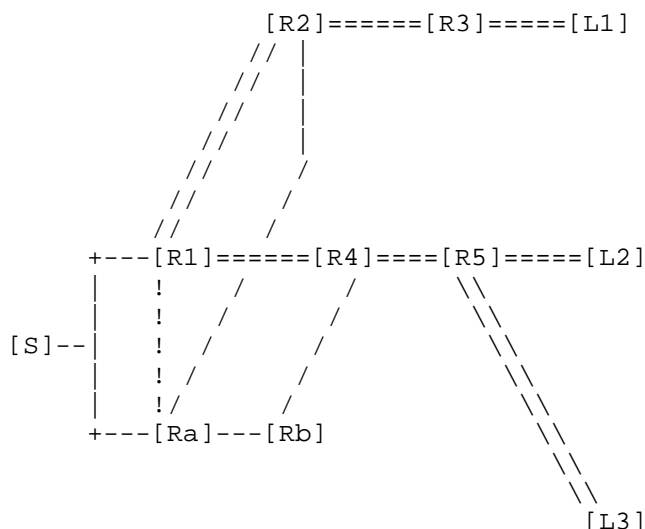


Figure 1: P2MP sub Tree for Locally Protecting Ingress

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. Protocol Extensions

This section describes a few of ways to extend the existing protocols for supporting TE LSP ingress local protection. Three approaches are discussed. The first one mainly uses a couple of new RSVP-TE messages. The second one adds some new objects into existing RSVP-TE messages. The third one mainly uses OSPF opaque LSAs.

6.1. New RSVP-TE Messages

This sub section presents two types of messages: LSP information message and LSP information confirmation 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.

LSP information confirmation messages are used to confirm that the corresponding LSP information messages are received. 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.

6.1.1. LSP Information Message

6.1.1.1. 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).

The <EXPLICIT_ROUTE> and <S2L sub-LSP descriptor list> contains the path from the backup ingress node to the next hops of the primary ingress, and then to the egresses. If the path from the backup ingress node to the next hops of the primary ingress is loose, the detailed path from the backup ingress node to the next hops needs to be computed.

The <RECORD_ROUTE> and <S2L sub LSP flow descriptor list> comprises the information about the path that the LSP traversed.

6.1.1.2. 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, provides and maintains local protection for the primary ingress according to the information in the information message.

6.1.2. Backup LSP for One-to-One Backup

When the backup ingress receives the LSP information message with the request for protection via the one-to-one backup method from the primary ingress, it 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.

6.1.2.1. 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 if the path from the backup ingress to the next hops is loose. 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.

6.1.3. Backup LSP for Facility Backup

The backup ingress selects or creates a backup P2MP LSP tunnel from itself to the next hop nodes of the primary LSP when it receives the LSP information message with a request for protection via the Facility backup method from the primary ingress.

If there exists a backup P2MP LSP tunnel from the backup ingress to the next hop nodes of the P2MP LSP that satisfies the constraints

given in the information message from the (primary) ingress, then this tunnel is selected; otherwise, a new backup P2MP LSP tunnel from the backup ingress to the next hop nodes of the P2MP LSP will be created.

After having a backup P2MP LSP tunnel, the backup ingress assigns an inner label (or upstream label) using upstream label assignment procedures for the primary LSP.

To signal the backup P2MP LSP, a backup LSP's PATH message is sent to each of the next hop nodes of the primary ingress of the protected LSP. This PATH message MUST include an Upstream Assigned Label object carrying the upstream label and an RSVP-TE P2MP LSP TLV within an IF_ID RSVP object, carrying the session object of the P2MP Bypass tunnel.

When the backup ingress detects a failure in the primary ingress of the protected P2MP LSP, it has to import the traffic for the protected P2MP LSP into the backup P2MP bypass tunnels using the upstream label assigned for this protected P2MP LSP as an inner label. The backup ingress MUST send PATH messages for the protected P2MP LSP.

6.1.4. LSP Information Confirmation Message

6.1.4.1. 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> <RRO>
    <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).

6.1.4.2. 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. If the backup LSP for locally protecting the primary ingress is available, the backup ingress node sets "local protection available" flag in the IPv4 (or IPv6) address sub-object of the RRO for the primary ingress and SHOULD send the updated confirmation message to the primary ingress.

The backup ingress node sets the "node protection" flag if the backup path protects against the failure of the primary ingress node, and, if the path does not, it clear the "node protection" flag.

The backup ingress node sets "bandwidth protection" flag if the backup path offers a bandwidth guarantee, and, if the path does not, it clear the "bandwidth protection" flag.

6.2. New RSVP-TE Objects

A desire for creating a backup LSP to locally protect the (primary) ingress of a P2MP LSP can be sent to a backup ingress from the primary ingress in a PATH message, which comprises the information about the P2MP LSP and the desire.

6.2.1. Information about Existing LSP

There are <style> and <flow descriptor list> normally in a RSVP-TE RESV message. They are "new" to a PATH message. The primary ingress of the P2MP LSP MAY add them into the PATH message to be sent to the backup ingress for locally protecting the (primary) ingress after it receives a RESV message.

<style> and <flow descriptor list> contains the information about the path that the LSP traverses. In fact, we may just add <RECORD_ROUTE> and <S2L sub LSP flow descriptor list> into the PATH message instead of <style> and <flow descriptor list>.

The primary ingress MUST send a updated PATH message to the backup ingress whenever there is a change in the message. It MAY send the same message to the backup ingress every refresh interval if there is no change.

6.2.2. Desire for Locally Protecting Ingress

A desire for locally protecting the (primary) ingress of a P2MP LSP MAY be implied by the "new" objects in the PATH message sent from the

primary ingress to the backup ingress.

It would be better to explicitly indicate the desire in the PATH message through using a new flag or new object.

The (primary) ingress of the LSP MAY request Ingress Local Protection by setting a bit in the Attributes Flags TLV. It is RECOMMENDED to use the LSP_REQUIRED_ATTRIBUTES object for the TLV.

A backup ingress that supports the Attributes Flags TLV and recognizes this bit MUST support Ingress Local Protection.

6.2.3. Backup LSP for One-to-One Backup

When the backup ingress receives the PATH message with the request for Ingress Local Protection and the request for protection via the one-to-one backup method from the primary ingress, it stores the information in the message, constructs a PATH message for a backup LSP, and sends the PATH message downstream accordingly. If it has not received any PATH message from the primary ingress 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, constructs a PathTear message, and send the PathTear message downstream accordingly.

The PATH message constructed for the backup LSP contains an EXPLICIT_ROUTE object and the objects in the S2L sub-LSP descriptor list. These objects represent a path from the backup ingress to the next-hop nodes of the primary ingress, and to the destination nodes of the P2MP LSP. The backup path from the backup ingress to the next-hop nodes of the primary ingress may be computed by the backup ingress. The path segment from the next-hop nodes of the primary ingress to the destination nodes of the P2MP LSP may be from the RECORD_ROUTE object and the objects for the S2L sub-LSP flow descriptor list in the PATH message received from the primary ingress.

6.2.4. Backup LSP for Facility Backup

The backup ingress selects or creates a backup P2MP LSP tunnel from itself to the next hop nodes of the primary LSP when it receives a PATH message with a request for Ingress Local Protection and a request for protection via the Facility backup method from the primary ingress.

If there exists a backup P2MP LSP tunnel from the backup ingress to the next hop nodes of the P2MP LSP that satisfies the constraints given in the PATH message from the (primary) ingress, then this

tunnel is selected; otherwise, a new backup P2MP LSP tunnel from the backup ingress to the next hop nodes of the P2MP LSP will be created.

After having a backup P2MP LSP tunnel, the backup ingress assigns an inner label (or upstream label) using upstream label assignment procedures for the primary LSP.

To signal the backup P2MP LSP, a backup LSP's PATH message is sent to each of the next hop nodes of the primary ingress of the protected LSP. This PATH message MUST include an Upstream Assigned Label object carrying the upstream label and an RSVP-TE P2MP LSP TLV within an IF_ID RSVP object, carrying the session object of the P2MP Bypass tunnel.

When the backup ingress detects a failure in the primary ingress of the protected P2MP LSP, it has to import the traffic for the protected P2MP LSP into the backup P2MP bypass tunnels using the upstream label assigned for this protected P2MP LSP as an inner label. The backup ingress MUST send PATH messages for the protected P2MP LSP.

6.3. OSPF Opaque LSA

The information about a P2MP LSP may be transferred through using an OSPF Opaque LSA.

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.

6.4. Mapping Traffic to Backup LSP

After the backup ingress node establishes a backup P2MP sub tree for protecting the primary ingress node of the P2MP LSP successfully, it may map a given set of traffic to the backup P2MP sub tree through creating a inactive forwarding entry with a FEC. This FEC indicates the traffic that the P2MP LSP carries.

In a normal operation, this inactive forwarding entry is not used to forward any data traffic. When the primary ingress node fails, this inactive forwarding entry is changed to active. Thus, the traffic

for the P2MP LSP is imported into the backup P2MP sub tree, and then merged into the P2MP LSP to be transported to the destinations.

The FEC can be configured on the backup ingress node. It can also be transferred to the backup ingress from the primary ingress through a FEC object or sub object.

The primary ingress node may send the FEC through one of the three options mentioned above to the backup ingress node, which creates a inactive forwarding entry with the FEC associated with the backup P2MP sub tree.

7. IANA Considerations

TBD

8. Acknowledgement

The authors would like to thank Richard Li, Rahul Aggarwal, Olufemi Komolafe, Rob Rennison, Neil Harrison, Kannan Sampath, Yimin Shen, Ronhazli Adam and Quintin Zhao for their valuable comments and suggestions on this draft.

9. References

9.1. Normative References

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

9.2. 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: huaimo.chen@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
UC Davis
USA

Email: liulei.kddi@gmail.com

Routing Area Working Group
Internet-Draft
Intended status: Standards Track
Expires: August 24, 2013

H. Gredler
Juniper Networks, Inc.
February 20, 2013

Advertising MPLS labels in IGPs
draft-gredler-rtgwg-igp-label-advertisement-02

Abstract

Historically MPLS label distribution was driven by session oriented protocols. In order to obtain a particular routers label binding for a given destination FEC one needs to have first an established session with that node.

This document describes a mechanism to distribute FEC/label mappings through flooding protocols. Flooding protocols publish their objects for an unknown set of receivers, therefore one can efficiently scale label distribution for use cases where the receiver of label information is not directly connected.

Application of this technique are found in the field of backup (LFA) computation, Label switched path stitching, traffic engineering and egress ASBR link selection.

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 August 24, 2013.

Copyright Notice

Copyright (c) 2013 IETF Trust and the persons identified as the document authors. All rights reserved.

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

Table of Contents

| | |
|---|---|
| 1. Introduction | 3 |
| 2. Motivation and Applicability | 3 |
| 2.1. Explicit One hop tunnels | 4 |
| 2.2. Egress ASBR Link Selection | 5 |
| 2.3. Explicit Path Routing through Label Stacking | 6 |
| 2.4. Stitching MPLS Label Switched Path Segments | 7 |
| 3. Acknowledgements | 8 |
| 4. IANA Considerations | 8 |
| 5. Security Considerations | 8 |
| 6. References | 9 |
| 6.1. Normative References | 9 |
| 6.2. Informative References | 9 |
| 6.3. References | 9 |
| Author's Address | 9 |

1. Introduction

MPLS label allocations are predominantly distributed by using the LDP [RFC5036], RSVP [RFC5151] or labeled BGP [RFC3107] protocol. All of those protocols have in common that they are session oriented, which means that in order to learn the Label Information database of a particular router one needs to have a direct control-plane session using the given protocol.

There are a couple of interesting use cases where the consumer of a MPLS label allocation may not be adjacent to the router having allocated the label. Bringing up an explicit session using existing label distribution protocols between the non-adjacent label allocator and the label consumer is the existing remedy for this dilemma.

For LDP protection routing LDP next next hop labels [NNHOP] have been proposed to provide the 2 hop neighborhood labels. While the 2 hop neighborhood provides good backup coverage for the typical network operator topology it is inadequate for some sparse for example ring like topologies.

Depending on the application, retrieval and setup of forwarding state of such >1 hop label allocations may only be transient. As such configuring and un-configuring the explicit session is an operational burden and therefore should be avoided.

2. Motivation and Applicability

It may not be immediate obvious, however introduction of Remote LFA [I-D.ietf-rtgwg-remote-lfa] technology has implied important changes for an IGP implementation. Previously the IGP had a one-way communication path with the LDP module. The IGP supplies tracking routes and LDP selects the best neighbor based upon FEC to tracking routes exact matching results. Remote LFA changes that relationship such that there is a bi-directional communication path between the IGP and LDP. Now the IGP needs to learn about if a label switched path to a given destination prefix has been established and what the ingress label for getting there is. The IGP needs to push that label for the tracking routes of destinations beyond a remote LFA neighbor.

Since the IGP now creates forwarding state based on label information it may make sense to distribute label by the IGP as well. This section lists example applications of IGP distribution of MPLS labels.

2.1. Explicit One hop tunnels

Deployment of Loop free alternate backup technology RFC 5286 [RFC5286] results in backup graphs whose coverage is highly dependent on the underlying Layer-3 topology. Typical network deployments provide backup coverage less than 100 percent (see RFC 6571 Section 4.3 for Results [RFC6571]) for IGP destination prefixes.

By closer examining the coverage gaps from the referenced production network topologies, it becomes obvious that most topologies lacking backup coverage are close to ring shaped topologies (Figure 1).

Remote LFA [I-D.ietf-rtgwg-remote-lfa] has introduced the notion of a "remote" LFA neighbor. This helper router which is both in P and Q space could forward the traffic to the final destination. Router 'H' is in P space, however due to the actual metric allocation router 'H' is not in Q space.

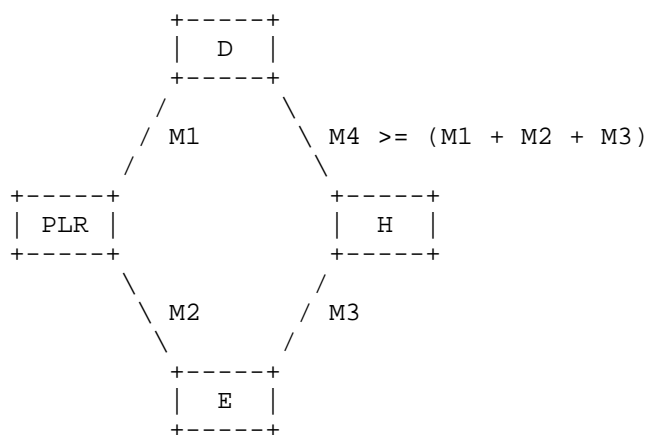


Figure 1: Coverage gap analysis

The protection router (PLR) evaluates for a primary path to destination 'D' if $\{E \rightarrow H \rightarrow D\}$ is a viable backup path. Because the metric $M4 \{H \rightarrow D\}$ is higher than the sum of the original primary path and the path from router 'H' to the PLR, this particular path would result in a loop and therefore is rejected.

Now consider that router 'H' would advertise a label for FEC 'D', which has the semantics that H will POP the label and forward to the destination node 'D'. This is done irrespective of the underlying IGP metric 'M4' it is a 'strict forwarding' label. The PLR router can now construct a label stack where the outermost label provides transport to router 'H'. The next label on the MPLS stack is the IGP

learned 'strict forwarding label' label. Note that the label 'strict forwarding' semantics are similar to a 1-hop ERO (Explicit route object). The Remote 'LFA' calculation would need to get changed, such that even if a node is not in PQ space, but rather in P space, it may get used as a backup neighbor if it advertises a strict forwarding label to the final destination. A recursive version of the algorithm is applicable as well as long a node in P space has some non looping LSP path to the final destination. The PLR router can now program a backup path irrespective of the undesirable underlying layer-3 topology.

Using existing tunnels for backup routing has been previously described in [I-D.bryant-ipfrr-tunnels]. Section 5.2.3 'Directed forwarding' describes an option to insert a single MPLS label between the tunnel and the payload. Traffic may thereby be directed to a particular neighbor. The mechanism described in this document, is an MPLS specific manifestation of 'Directed forwarding'.

2.2. Egress ASBR Link Selection

In the topology described in Figure 2. router 'S' is facing a dilemma. Router S receives a BGP route from all of its 4 upstream routers. Using existing mechanism the provider owning AS1 can control the loading of its direct links *to* its ASBR1 and ASBR2, however it cannot control the load of the links beyond the ASBRs, except manually tweaking the eBGP import policy and filtering out a certain prefix. It would be more desirable to have visibility of all four BGP paths and be able to control the loading of those four paths using Weighted ECMP. Note that the computation of the 'Weight' percentage and the component doing this computation (Router embedded or SDN) is outside the scope of this document.

If all the ASes would be under one common administrative control then the network operator could deploy a forwarding hierarchy by using [RFC3107] to learn about the remote-AS BGP nexthop addresses and associated labels. An ingress router 'S' would then stack the transport label to its local egress ASBR and the remote ASBR supplied label. In reality it is hard to convince a peering AS to deploy another protocol just in order to easier control the egress load on the WAN links for the ingress AS.

A 'strict forwarding' paradigm would solve this problem: An Egress ASBR (e.g. ASBR 1 and 2) allocates a strict forwarding label toward all of its peering ASes and advertises it into its local IGP. The forwarding state of all those labels is to POP off the label and forward to the respective interface. The ingress router 'S' then builds a MPLS label stack by combining its local transport label to ASBR3 or ASBR4 with the IGP learned label pointing to the remote-AS

ASBR.

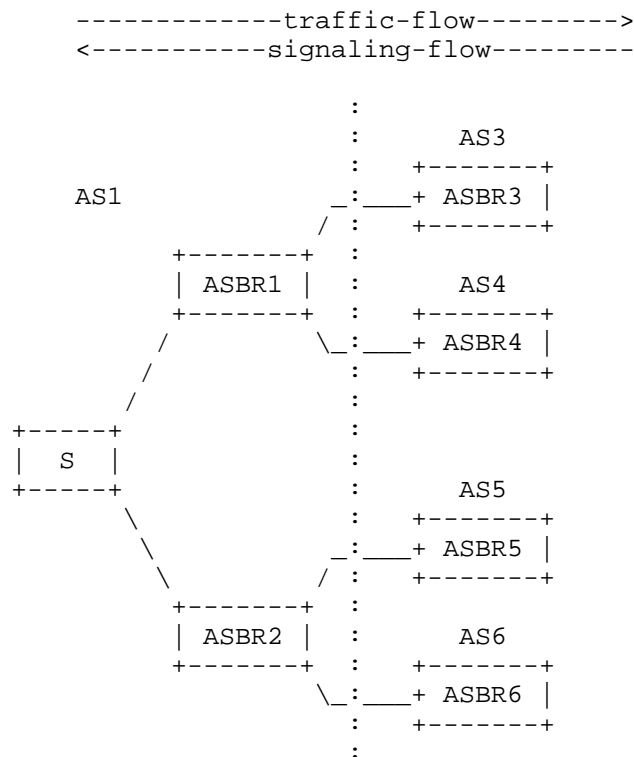


Figure 2: Egress ASBR Link selection

2.3. Explicit Path Routing through Label Stacking

IGP advertised strict forwarding labels can be utilized for constructing simple EROs via virtue of the MPLS label stack. In a classical traffic engineering problem (Figure 3) is illustrated. The best IGP path between {S,D} is {S, R3, R4, D}. Unfortunately this path is congested. It turns out that the links {S, R1}, {R1, R4} and {R2, R4} do have some spare capacity. In the past a C-SPF calculation would have passed the ERO {S, R1, R4, R2, D} down to RSVP for signaling. The conceptional problem with RSVP signaled paths is that they cannot be shared with other nodes in the network. Hence all potential ingress routers need to setup their "private" ERO path and allocate network signaling resources and forwarding state.

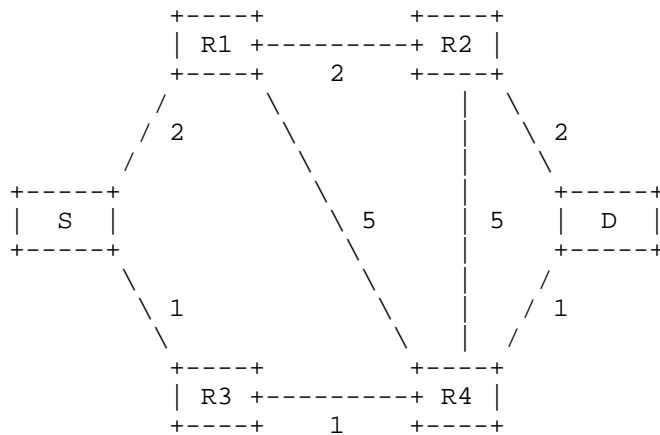


Figure 3: Explicit Routing using Label stacking

Consider now every router along the path does advertise a strict forwarding label for its direct neighbor. Router S could now construct a couple of paths for avoiding the hot links without explicitly signaling them.

- o {S, R1, R2, D}
- o {S, R1, R4, D}
- o {S, R1, R4, R2, D}

Note that not every hop in the ERO needs to be unique label in the label stack. This is undesired as existing forwarding hardware technology has got upper limits how much labels can get pushed on the label stack. In fact an existing tunnel (for example LDP tunnel {S, R1, R2}) can be reused for certain path segments.

2.4. Stitching MPLS Label Switched Path Segments

One of the shortcomings of existing traffic-engineering solutions is that existing label switched paths cannot get advertised and shared by many ingress routers in the network. In the example network (Figure 4) a LSP with an ERO of {R4, R2, R6} has been established in order to utilize two unused north / south links. The only way to attract traffic to that LSP is to advertise the LSP as a forwarding adjacency. This causes loss of the original path information which might be interesting for a potential router which might want to use this LSP for backup purposes. A computing router would need to have all underlying fate-sharing and bandwidth utilization information.

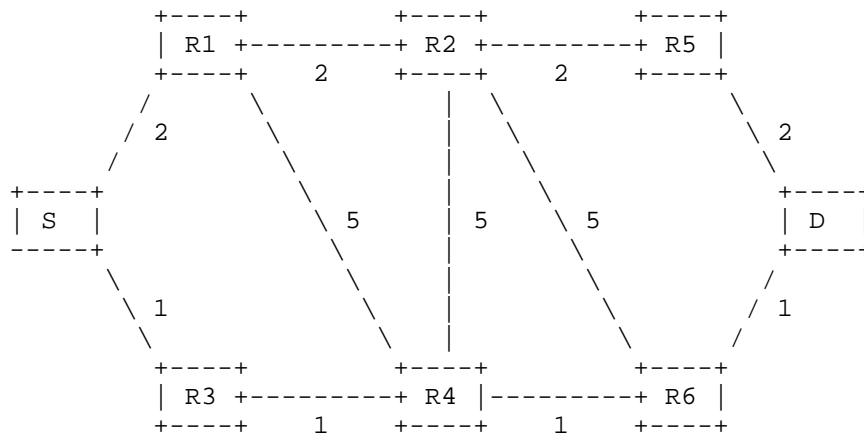


Figure 4: Advertising path segments

The IGP on R4 can now advertise the LSP segment by advertising its ingress label and optionally pass the original ERO, such that any upstream router can do their fate-sharing computations. Potential ingress routers now can use this LSP as a segment of the overall LSP. Furthermore ingress routers can combine label advertisements from different routers along the path. For example router S could stack its LDP path to R2 {S, R1, R2} plus the IGP learned RSVP LSP {R4, R5, R6} plus a strict forwarding label {R6, D}.

3. Acknowledgements

Many thanks to Yakov Rehkter, Ina Minei, Stephane Likowski and Bruno Decraene for their useful comments.

4. IANA Considerations

This memo includes no request to IANA.

5. Security Considerations

This document does not introduce any change in terms of IGP security. It simply proposes to flood existing information gathered from other protocols via the IGP.

6. References

6.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC3107] Rekhter, Y. and E. Rosen, "Carrying Label Information in BGP-4", RFC 3107, May 2001.
- [RFC5036] Andersson, L., Minei, I., and B. Thomas, "LDP Specification", RFC 5036, October 2007.
- [RFC5151] Farrel, A., Ayyangar, A., and JP. Vasseur, "Inter-Domain MPLS and GMPLS Traffic Engineering -- Resource Reservation Protocol-Traffic Engineering (RSVP-TE) Extensions", RFC 5151, February 2008.
- [RFC5286] Atlas, A. and A. Zinin, "Basic Specification for IP Fast Reroute: Loop-Free Alternates", RFC 5286, September 2008.
- [RFC6571] Filsfils, C., Francois, P., Shand, M., Decraene, B., Uttaro, J., Leymann, N., and M. Horneffer, "Loop-Free Alternate (LFA) Applicability in Service Provider (SP) Networks", RFC 6571, June 2012.

6.2. Informative References

- [I-D.bryant-ipfrr-tunnels]
Bryant, S., Filsfils, C., Previdi, S., and M. Shand, "IP Fast Reroute using tunnels", draft-bryant-ipfrr-tunnels-03 (work in progress), November 2007.
- [I-D.ietf-rtgwg-remote-lfa]
Bryant, S., Filsfils, C., Previdi, S., Shand, M., and S. Ning, "Remote LFA FRR", draft-ietf-rtgwg-remote-lfa-01 (work in progress), December 2012.

6.3. References

- [NNHOP] Chen, E., Shen, N., and A. Tian, "Discovering LDP Next-Nexthop Labels", November 2005, <<http://tools.ietf.org/html/draft-shen-mpls-ldp-nnhop-label-02>>.

Author's Address

Hannes Gredler
Juniper Networks, Inc.
1194 N. Mathilda Ave.
Sunnyvale, CA 94089
US

Email: hannes@juniper.net

MPLS Working Group
Internet Draft

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

Intended status: Informational

Expires: August 24, 2013

February 25, 2013

Framework for Point-to-Multipoint MPLS-TP OAM
draft-hmk-mpls-tp-p2mp-oam-framework-02.txt

Status of this Memo

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

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

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

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

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

This Internet-Draft will expire on August 24, 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 discusses and specifies the P2MP framework primarily related to OAM and related management in MPLS-TP networks. This document mainly refers to RFC5654 and RFC6371. The main focus is on the details that are not covered or not clarified in relevant RFCs such as RFC5654, RFC5860, RFC5921, RFC5951, RFC6371, and draft-mpls-tp-p2mp-framework.

Note: This I-D was made and updated including the discussions in ITU-T SG15, which were described in Liaison Statements such as (<https://datatracker.ietf.org/liaison/1235/>)

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 | 5 |
| 3.1. OAM functions for proactive monitoring | 8 |
| 3.1.1. Continuity Check and Connectivity Verification..... | 11 |
| 3.1.2. Remote Defect Indication | 11 |
| 3.1.3. Alarm Reporting | 12 |

| | |
|---|----|
| 3.1.4. Lock Reporting | 12 |
| 3.1.5. Packet Loss Measurement | 12 |
| 3.1.6. Packet Delay Measurement | 12 |
| 3.1.7. Client Failure Indication | 12 |
| 3.2. OAM functions for on-demand monitoring | 12 |
| 3.2.1. Connectivity verification | 12 |
| 3.2.2. Packet loss measurement | 13 |
| 3.2.3. Diagnostic tests | 13 |
| 3.2.4. Route Tracing | 13 |
| 3.2.5. Packet delay measurement | 13 |
| 3.3. OAM functions for administration control | 13 |
| 3.3.1. Lock Instruct | 13 |
| 4. Security Considerations | 13 |
| 5. IANA Considerations | 13 |
| 6. References | 14 |
| 6.1. Normative References | 14 |
| 6.2. Informative References | 14 |
| 7. Acknowledgments | 14 |

1. Introduction

The demand for P2MP traffic is expected to quickly increase due to the increase in new services such as IP-TV, compressed & uncompressed video distribution, and smart TV. In light of the global trend in improving energy efficiency as well as general network cost reduction, a point-to-multipoint (P2MP) transport function in MPLS-TP could be one of the solutions for providing these services from the perspective of efficient use of network resources.

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

- Traffic-engineered point-to-multipoint (P2MP) transport paths.(item 6).
- Unidirectional point-to-multipoint(P2MP) 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 the 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. With a unidirectional P2MP transport path, two cases are assumed as per Section 3.3 of RFC6371[3]. One is when no return path exists or not used and the other is when an "out-of-band" return path exists and used.

In I-D[4], only a summary of various items specific to MPLS-TP P2MP framework. For example, according to the editor's note, this section will contain a summary of P2MP OAM, as described in RFC6371 [3], which defines the overall OAM architecture for MPLS-TP.

Therefore, this draft intends to specify details of a P2MP framework that complements P2MP requirements and the framework of existing RFCs, particularly in terms of OAM, management, and recovery.

Note: MPLS-TP functions that are applicable specifically to P2MP transport paths are outside the scope of RFC5921.

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

EMS Element management system

LSP Label Switched Path

NE Network Element

NMS Network Management System

2.2. Definitions

None

3. P2MP OAM and management

3.1. General aspects of architecture

3.1.1. Return path

The support of P2MP OAM on the data path should be independent of the availability of a return path or the mechanism that supports the return path. Basically, only unidirectional P2MP is supported in MPLS-TP. This means that an "in-band" return path is out of the scope of MPLS-TP requirements. In this section, two cases, with out-band return path and without return path, are considered basic and the requirements that should be met when return paths exist should be independently specified in other document, if needed.

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, most OAM requirements in RFC5860 can be met by supporting the management interface through which EMS/NMS can retrieve the received OAM packets.

The "return path" may be considered to be directed to the entity that originally requested the measurements because this may not be the head end of the P2MP connection. Therefore, the following return path should be distinctly differentiated.

RP-N: A return path to the EMS/NMS through the management interface (RP-N) (this case is referred to as that in which no return path exists)

RP-HE: A return path to a head end (root) of a P2MP path using any kind of out-of-band path (this case is referred to as that in which an out-of-band return path exists)

The interpretation of return path usually corresponds to RP-HE. These two kinds of return paths may be applied at the same time, depending on the situations.

3.1.2. M-leaves management scenario in P2MP path

Generally, a function to monitor only the subset leaves of a P2MP transport path is required to appropriately monitor the status of P2MP transport paths. The supplemental requirements are as follows.

- 1) M-leaves management, which enables NMS to perform OAM functions at a set of leaves on a P2MP transport path, must be supported.

- 2) M-leaves must be selectable by the operator or administrator using NMS.
- 3) M-leaves management should be independently enabled/disabled in each OAM function.
- 4) In M-leave monitoring, one scenario should be selected to avoid future interoperability problems between related entities (NE, EMS, and NMS).

There are four scenarios considered in MPLS-TP networks that consist of NEs, EMS, and NMS.

In scenario 1, OAM protocol extension is necessary. OAM packets sent from the source MEP must include a subset of leaf-MEPs. A sink MEP determines if it should be notified of the management process within an NE based on the leaf-IDs included in the OAM packet. However, this is not supported in RFC6371.

In scenario 2, OAM packets that are supported in RFC6371 and are targeted at all leaves can be utilized. As a result, no extension is necessary in the P2MP OAM protocol. On the other hand, a subset of M-leave/sink MEPs must be configured at an EMS from an NMS. In addition, a pre-configuration of a subset of M-leave/sink MEPs is needed at related NEs from the EMS. Only the notification-enabled M-leaves/nodes notify the EMS of its monitoring results.

In scenario 3, OAM packets that are supported in RFC6371 and are targeted at all leaves can also be utilized. There is no P2MP OAM protocol extension. On the other hand, NMS configuration on M-leaves/sink MEPs is needed. In addition, a subset of M-leave/sink MEPs must be configured at the EMS from the NMS. However, no pre-configuration of a subset of M-leaves/NEs is needed.

In scenario 4, OAM packets that are supported in RFC6371 and are targeted at all leaves can also be utilized. There is no P2MP OAM protocol extension. Only NMS configuration on M-leaves/sink MEPs is needed. A configuration of a subset of M-leave/sink MEPs at the EMS from the NMS is not necessary. No pre-configuration of a subset of M-leaves/NEs is needed.

Considering some negative impacts such as the efficient use of a data communication network (DCN), insufficient manageability of network element (NE), traffic congestion at EMS/NMS, and heavy load for OAM packet processes at EMS/NMS, scenario 2 is required in MPLS-TP p2mp network.

3.1.3. Refinement of existing requirements on P2MP transport path

MPLS-TP RFCs are sufficiently mature in terms of the requirements and framework of MPLS-TP P2P. On the other hand, in terms of MPLS-TP P2MP, some parts of MPLS-TP RFCs and Recommendations could be refined and clarified.

(R1) CV requirement of RFC5860

CV is ambiguously defined in RFC5860 "MPLS-TP OAM requirement". According to this definition of RFC5860, it seems to be source-MEP oriented and not correct in P2MP.

Current text: The MPLS-TP OAM toolset MUST provide a function to enable an End Point to determine whether or not it is connected to specific End Point(s) by means of the expected PW, LSP, or Section.

In unidirectional P2MP, the source MEP cannot determine whether or not it is connected to specific End Point(s). Therefore, in P2MP, the definition of connectivity verification should be corrected in P2MP framework draft and OAM Recommendation as follows.

Proposed text: The MPLS-TP OAM toolset MUST provide a function to enable a sink End Point to determine whether or not it is connected to a specific source End Point by means of the expected PW or LSP.

(R2) CC Requirement of RFC6371

According to RFC6371, it is assumed that CC means that CC OAM packet does not include either a source MEP or destination MEP. Only unidirectional P2MP is supported in MPLS-TP, so the continuity of the CC OAM packets are received by sink MEPs, and a sink MEP should notify the equipment fault management process of the detected defect. However, the following current text doesn't correctly describe the unidirectional feature that is specific to P2MP transport path. Therefore, the requirement should be modified.

Current text in RFC: Proactive Continuity Check functions, as required in Section 2.2.2 of RFC 5860 [11], are used to detect a loss of continuity (LOC) defect between two MEPs in an MEG. Proactive Connectivity Verification functions, as required in Section 2.2.3 of RFC 5860 [11], are used to detect an unexpected connectivity defect between two MEGs (e.g., mismerging or misconnection), as well as unexpected connectivity within the MEG with an unexpected MEP.

Proposed text: Proactive Continuity Check functions, as required in Section 2.2.2 of RFC5860, are used to detect a loss of continuity

(LOC) defect from the source MEP to sink MEP(s). Proactive Connectivity Verification functions, as required in Section 2.2.3 of RFC5860, are used to detect an unexpected connectivity defect from the source MEP to sink MEP(s) (e.g., mismerging or misconnection), as well as unexpected connectivity within MEG with an unexpected source MEP.

(R3) Optional requirements on CC-V OAM packets

In a P2MP transport path, it is highly desirable that in order to save OAM bandwidth consumption, CV, when used, be linked with CC into CC-V OAM packets.

3.1.4. Addition and removal of branch tree in P2MP transport path

When additional branches, in other words, additional destination NEs (leaves) need to be added to an existing transport path after a connection service is provided via the P2MP path, an operator must be capable of adding a new branch tree to the P2MP transport path flexibly from any point on the path without service interruption. The reason is that merging and crossover of the P2MP LSP branch tree must be rejected because it is not efficient in terms of network resources. As a result, the following requirement must be supported in the MPLS-TP P2MP transport path.

3.2. General aspects of P2MP OAM

P2MP transport paths are unidirectional; therefore, there is generally no in-band return path as in the MPLS-TP transport path per se. However, there are basically two approaches for handling OAM requirements in P2MP MPLS-TP.

The first one is used to report the results of the monitoring/measurement of OAM packets from the OAM target node to the EMS/NMS when the NMS usually instantiates OAM functions and requires the results of OAM monitoring functions. This approach is called RP-N. The second approach is the return path to a root (source MEP) of a P2MP path using different methods such as a unidirectional p2p transport paths, and other technology-layers, such as IP, Ethernet, and OTN, when an NE within which a root MEP resides instantiates OAM functions or receive results of OAM monitoring functions. This approach is called as RP-HE. The following requirements are supported in terms of network elements when considering RP-N.

1. OAM functions of a MEG of a P2MP transport path should be configurable using the EMS/NMS.

2. Source nodes at which the source MEP reside and OAM packets are generated should receive OAM related information such as enabling/disabling OAM functions and setting/changing OAM attributes from the EMS/NMS on a P2MP transport path.
3. Sink nodes at which targeting MIPs or MEPs reside and OAM packets are parsed should report OAM related information such as OAM monitoring results and consequent OAM actions to the EMS/NMS.
4. Each OAM function of a P2MP transport path should be able to be independently configured using the EMS/NMS based on the classification of OAM functional requirements in RFC5860.
5. An on-demand OAM function must be able to perform an OAM function for only a specific target MIP or MEP as well as all MEPs in a P2MP transport path, as specified in Section 3.7 of RFC6371[3].
6. To manage M leaves(i.e., subset of all leaves) in an on-demand OAM function from the EMS/NMS, a unified mechanism must be provided.

Note: Currently, sending an OAM packet that is targeted at a subset of M leaves by using an aggregating mechanism such as an OAM packet including several MIP or MEP identifiers is out of the scope of RFC6371[3] as described in Section 3.7 of that document.

7. Mismatches of configuration information between a root MEP and any leaf-MEP, at which proactive or on-demand monitoring is enabled, should be detected as a configuration mismatch alarm and be reported to the EMS/NMS by parsing received OAM packets, particularly when a static setting is applied.

Generally when each OAM function is enabled, as described in Section 5.1 of RFC6371[3], the source MEP function should be enabled prior to the corresponding sink MEPs' function.

Regarding configuration considerations, the following are additional requirements for unidirectional P2MP transport path, particularly when RP-HE does not exist.

8. The configuration of each OAM function between the source MEP and sink MEP(s) in an MEG of a transport path should be able to be synchronized using the NMS, when a new P2MP transport path is set.
9. OAM functions of a newly added/deleted branch transport path from any point of an existing transport path must be able to be configured and enabled/disabled on a newly integrated/combined P2MP transport path without affecting client traffic to existing

end points of the P2MP transport path other than the added/removed branch transport path.

10. The configuration of newly added/removed specific sink MEP(s) to the existing source MEP in the MEG in proactive monitoring should be able to be synchronized with that of the source MEP by using the NMS.
 11. The EMS/NMS should provide a tool for manually configuring consistent values of each piece of configuration information to a root MEP and all the related leaf MEPs in a MEG of a P2MP transport path for both pro-active and on-demand OAM functions.
 12. Mismatches of configuration information between a leaf MEP and any other leaf MEP(s) or a root MEP and leaf MEP(s), at which proactive monitoring will be enabled, should be able to be detected through the configuration management process of the EMS/NMS as a configuration mismatch alarm or notification without receiving OAM packets from a source MEP (before OAM functions are enabled).
- Note: This requirement is not necessary if the EMS/NMS provides a tool to manually configure a consistent value of each piece of configuration information to a root MEP.
13. The enabling or disabling of proactive OAM functions and configuration mismatch alarms of the OAM functions must be independently configurable at each leaf-MEP as well as on all the leaf MEPs on a P2MP transport path, considering maintenances or a case in which one or more leaf MEPs is newly added or removed later.
 14. Mismatches of configuration information between a leaf MEP and any other leaf MEP(s) or a root MEP and leaf MEP(s), at which on-demand OAM monitoring is enabled, must be detected as a configuration management process before conducting OAM functions.

3.3. OAM functions for proactive monitoring

The proactive OAM functions are used to detect a fault/defect or to automatically reports a change in the status of a transport path.

3.3.1. Continuity Check and Connectivity Verification(CC-V)

The continuity Check function enables one or more leaf MEPs on a unidirectional P2MP transport path to monitor the continuity of OAM packets from root MEP and detect one or more loss of continuity(LOC) defects between the root MEP and leaf MEPs.

The connectivity verification function enables one or more leaf MEPs on a 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)

As described in Sections 2.2.2 and 2.2.3 of RFC5860[2], CC-V MUST be supported even when RP-HE does not exist.

As described in RFC6371[3], CC-V OAM packets are used for a P2MP transport path. Defect detection mechanisms in P2MP transport paths are the same as those of the P2MP transport path specified in section 5.1.1 of RFC6371 [3]. That is, loss of continuity(LoC) defect, mis-connectivity defect, period mis-configuration defect and unexpected encapsulation defect. Entry and exit criteria are also the same as those of the P2MP transport paths in RFC6371 [3]. However, in a P2MP transport path, all the leaf MEPs that detect a defect must be identified and differentiated from a normal leaf MEP(s), which does not detect a defect.

Configuration is specified in Section 5.1.3 of RFC6371[3]. The following configuration information must be configured: MEG-ID, MEP-ID, list of the other MEPs in the MEG that are different between the root MEP and leaf MEP, PHB for E-LSP and transmission rate.

Consequent actions of a unidirectional P2MP transport path are also covered in Section 5.1.2 of RFC6371 [3]. Operators should be able to enable/disable each consequent action.

All MEPs inside a MEG need to be configured and retain the information when a proactive OAM function is enabled, as described in Section 5.1.3 of RFC6371[3]. If there is no RP-HE, it is premised that the EMS/NMS exists. Therefore, the above parameters are statically configured.

3.3.2. Remote Defect Indication

This OAM function is not available on a P2MP transport path when return paths do not exist. This OAM function can be implemented only

in RP-HE. However, the return path is out of the scope of MPLS-TP requirements.

3.3.3. Alarm Reporting

FFS

3.3.4. Lock Reporting

For further study(FFS)

3.3.5. Packet Loss Measurement

FFS

3.3.6. Packet Delay Measurement

FFS

3.3.7. Client Failure Indication

FFS

3.4. OAM functions for on-demand monitoring

3.4.1. Connectivity verification

The connectivity verification function enables one or more leaf MEPs on a 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)

1. Connectivity verification functions MUST be supported when return paths in a unidirectional P2MP transport path do not exist.

As described in RFC6371 [3], CC-V OAM packets are used for a P2MP transport path. Defect detection mechanisms in P2MP transport paths are the same as those of the 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 and exit criteria are also the same as those of the P2MP transport path in RFC6371 [3]. Moreover, consequent actions of a unidirectional P2MP transport path are also covered in Section 5.1.2 of the RFC [3]

Regarding configuration consideration, the following additional requirements on a unidirectional P2MP transport path when a return path does not exist.

3.4.2. Packet loss measurement

FFS

3.4.3. Diagnostic tests

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

Other requirements are ffs.

3.4.4. Route Tracing

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

Other requirements are ffs.

3.4.5. Packet delay measurement

FFS

3.5. OAM functions for administration control

3.5.1. Lock Instruct

FFS.

4. Security Considerations

This document does not 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 all, "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-mpls-tp-p2mp-framework-00, January 2013

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
Intended status: Standards Track
Expires: May 12, 2013

Zafar Ali
Rakesh Gandhi
Tarek Saad
Cisco Systems, Inc.
Robert H. Venator
Defense Information Systems Agency
Yuji Kamite
NTT Communications Corporation
November 8, 2012

Signaling RSVP-TE P2MP LSPs in an Inter-domain Environment
draft-ietf-mpls-inter-domain-p2mp-rsvp-te-lsp-00

Abstract

Point-to-MultiPoint (P2MP) Multiprotocol Label Switching (MPLS) and Generalized MPLS (GMPLS) Traffic Engineering Label Switched Paths (TE LSPs) are established using signaling procedures defined in [RFC4875]. However, [RFC4875] does not address several issues that arise when a P2MP-TE LSP is signaled in inter-domain networks. One such issue is the computation of a loosely routed inter-domain P2MP-TE LSP paths that are re-merge free. Another issue is the reoptimization of the inter-domain P2MP-TE LSP tree vs. an individual destination(s), since the loosely routing domain ingress border node is not aware of the reoptimization scope. This document defines the required protocol extensions needed for establishing and reoptimizing P2MP MPLS and GMPLS TE LSPs in inter-domain 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 May 12, 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. Summary of Solutions | 4 |
| 1.2. Path Computation Techniques | 5 |
| 1.3. Use cases | 5 |
| 2. Conventions used in this document | 6 |
| 3. Control Plane Solution For Re-merge Handling | 6 |
| 3.1. Single Border Node For All S2Ls | 6 |
| 3.2. Crankback and PathErr Signaling Procedure | 6 |
| 4. Data Plane Solution For Re-merge Handling | 8 |
| 4.1. P2MP-TE Re-merge Recording Request Flag | 8 |
| 4.2. P2MP-TE Re-merge Present Flag | 8 |
| 4.3. Signaling Procedure | 9 |
| 5. Intra-domain P2MP-TE LSP Re-merge Handling | 10 |
| 6. Reoptimization Handling | 11 |
| 6.1. P2MP-TE Tree Re-evaluation Request Flag | 11 |
| 6.2. Preferable P2MP-TE Tree Exists Flag | 11 |
| 6.3. Signaling Procedure | 11 |
| 7. Compatibility | 13 |
| 8. Security Considerations | 13 |
| 9. IANA Considerations | 13 |
| 10. Acknowledgments | 14 |
| 11. References | 15 |
| 11.1. Normative References | 15 |
| 11.2. Informative References | 15 |
| Author's Addresses | 16 |

1. Introduction

[RFC4875] describes procedures to set up Point-to-Multipoint (P2MP) Traffic Engineering Label Switched Paths (TE LSPs) for use in MultiProtocol Label Switching (MPLS) and Generalized MPLS (GMPLS) networks.

As with Point-to-Point (P2P) TE LSPs, P2MP TE LSP state is managed using RSVP messages. While the use of RSVP messages is mostly similar to their P2P counterpart, P2MP LSP state differs from P2P LSP in a number of ways. In particular, the P2MP LSP must also handle the "re-merge" problem described in [RFC4875] section 18.

The term "re-merge" refers to the situation when two source-to-leaf (S2L) sub-LSPs branch at some point in the P2MP tree, and then intersect again at a another node further downstream the tree. This may occur due to discrepancies in the routing algorithms used by different nodes, errors in path calculation or manual configuration, or network topology changes during the establishment of the P2MP LSP. Such re-merges are inefficient due to the unnecessary duplication of data and also consume additional network resources. Consequently one of the requirements for signaling P2MP LSPs is to choose a P2MP path that is re-merge free. In some deployments, it may also be required to signal P2MP-TE LSPs that are both re-merge and crossover free [RFC4875].

For the purposes of this document, a domain is considered to be any collection of network elements within a common sphere of address management or path computational responsibility. Examples of such domains include Interior Gateway Protocol (IGP) areas and Autonomous Systems (ASes). A border node is a node between different routing domains.

The re-merge free requirement becomes more acute to address when P2MP LSP spans multiple domains. This is because in an inter-domain environment, the ingress node may not have topological visibility into other domains to be able to compute and signal a re-merge free P2MP LSP. In that case, the border node for a new domain will be given loose next hops for one or more destinations in a P2MP LSP. A border node computes paths in its domain by individually expanding the loose next hops for the destinations when signaled to it. A border node can ensure that it computes the re-merge free paths while performing loose hop ERO expansions by individually grafting destinations. Note that the computed P2MP tree by a border node in this case may not be optimal. When processing a Path message, the border node may not have knowledge of all the destinations of the P2MP LSP; for example, in the case when not all S2L sub-LSPs pass through this border node. In that case, existing protocol mechanisms

do not provide sufficient information for it to be able to expand the loose hop(s) such that the overall P2MP LSP tree is guaranteed to be re-merge free.

[RFC4875] specifies two approaches to handle re-merge conditions. The first method is based on control plane handling the re-merge. In this case the node detecting the re-merge condition, i.e. the re-merge node initiates the removal of the re-merge sub-LSP(s) by sending a PathErr message(s) towards the ingress node. However, this can lead to a deadlock in setting up the P2MP LSP in certain cases; for example, when the first S2L setup causes the re-merge with all subsequent S2Ls in the tree. The second method is based on the data plane handling the re-merge condition. In this case, the re-merge node allows the re-merge condition to persist, but data from all but one incoming interface is dropped at the re-merge node. This ensures that duplicate data is not sent on any outgoing interface. However, network resources (such as bandwidth capacity) are wasted as long as re-merge condition persists which is inefficient.

[RFC4736] defines procedures and signaling extensions for reoptimizing an inter-domain P2P LSP. Specifically, an ingress node sends a "path re-evaluation request" to a border node by setting a flag (0x20) in SESSION_ATTRIBUTES object in a Path message. A border node sends a PathErr code 25 (notify error defined in [RFC3209]) with sub-code 6 to indicate "preferable path exists" to the ingress node. The ingress node upon receiving this PathErr may initiate reoptimization of the LSP. [RFC4736] however does not define a procedure to reoptimize the entire P2MP LSP as a whole tree.

As per [RFC4875] Section 14, for a P2MP LSP, an ingress node may reoptimize the entire P2MP LSP by resignaling all destinations (Section 14.1, "Make-before-Break") or may reoptimize individual the destinations (Section 14.2 "Sub-Group-Based Re-Optimization"). Generally speaking make-before-break is considered available for "whole" P2MP LSP reoptimization, but it can also be used for reoptimizing physical routes for specific sub-LSP(s). The Sub-Group-Based reoptimization is not always applicable because it can lead to data duplication inside the backbone.

1.1. Summary of Solutions

This document defines RSVP-TE signaling procedures for P2MP LSP to handle the re-merge condition when using either the control plane or data plane approach. The procedures are applicable to both MPLS TE and GMPLS networks.

The control plane solution for the re-merge problem makes use of the crankback signaling mechanism of the RSVP protocol. [RFC5151]

describes such mechanisms for applying crankback to inter-domain P2P LSPs, but does not cover P2MP LSPs. Also, crankback mechanisms for P2MP LSPs are not addressed by [RFC4875]. This document describes how crankback signaling extensions for MPLS and GMPLS RSVP-TE defined in [RFC4920] can be used for setting up P2MP TE LSPs to resolve re-merges.

The data plane solution for the re-merge problem described in [RFC4875] is extended by using a new flag in the LSP_ATTRIBUTES TLV (in a Path message) and a new flag in RRO Attributes Sub-object (in a Resv message) in RSVP. The LSP_ATTRIBUTES TLV (in a Path message) and RRO Attributes Sub-object (in a Resv message) have been defined in [RFC5420]. This document describes how these new flags can be used to handle P2MP re-merge conditions efficiently.

For P2MP LSP, a border node may have loosely routed entire or part of the P2MP LSP by expanding EROs in Path messages of the destinations. Border node does not know with the signaling procedure defined in [RFC4736] if an ingress node is requesting a reoptimization for an individual destination(s) or reoptimization of the entire P2MP tree. Signaling extension and procedure are defined in this document to handle reoptimization of an individual destination(s) and the reoptimization of the entire P2MP tree. Basically, a new query message is defined in LSP_ATTRIBUTES TLV to request for a "P2MP-TE Tree Re-evaluation" and a new sub-code is defined for PathErr message to indicate "Preferable P2MP-TE Tree Exists".

1.2. Path Computation Techniques

This document focuses on the case where the ingress node does not have full visibility of the topology of all domains and is therefore not able to compute the complete P2MP tree. Rather, it includes loose hops to traverse the domains for which it does not have full visibility and ingress border nodes(s) of each transit domain is responsible for expanding those loose hops.

The solution presented in this document do not guarantee optimization of the overall P2MP tree across all domains. Path Computation Element (PCE) can be used, instead, to address global optimization of the overall P2MP tree.

1.3. Use cases

Service providers having a network with multiple routing domains are interested to use the network for P2MP-TE LSPs. This allows the service providers to use the network to carry multicast and broadcast traffic (such as video). Service providers can deploy the VPLS and MVPN services in the network using inter-domain P2MP TE LSPs. The use

case is for P2MP TE LSPs across multiple routing domains that belong to a single administrative area. Use case for the Multiple administrative domains (e.g. autonomous systems) is outside the scope of this document.

2. Conventions used in this document

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

3. Control Plane Solution For Re-merge Handling

It is RECOMMENDED that boundary re-routing is requested for P2MP LSPs traversing multiple domains. This is because border nodes that are expanding loose hops are typically best placed to correct any re-merge errors that occur within their domain, not the ingress node.

3.1. Single Border Node For All S2Ls

It is RECOMMENDED that the ingress node of a P2MP LSP selects the same ingress border node in the loose hop ERO for all sibling S2L sub-LSPs that transit through a given domain. The reason is that it will increase the possibility of re-merge downstream if two or more border nodes have roles simultaneously to expand loose EROs. An ingress border node that performs the loose ERO expansion for individual sub-LSP(s) has the necessary state information for the destinations transiting through its domain to ensure computed P2MP tree is re-merge free.

3.2. Crankback and PathErr Signaling Procedure

As mentioned in [RFC4875], in order to avoid duplicate traffic, the re-merge node MAY initiate the removal of the re-merge S2L sub-LSPs by sending a PathErr message to the ingress node of the S2L sub-LSP.

Crankback procedures for rerouting around failures for P2P RSVP-TE LSPs are defined in [RFC4920]. These techniques can also be applied to P2MP LSPs to handle re-merge conditions, as described in this section.

If an ingress border node on the path of the P2MP LSP is unable to find a route that can supply the required resources or that is re-merge free, it MUST generate a PathErr message for the subset of the S2L sub-LSPs which it is not able to route. For this purpose the ingress border node SHOULD try to find a minimum subset of S2L sub-LSPs for which the PathErr needs to be generated towards the ingress

node. These are the S2L sub-LSPs on an incoming interface that has less number of S2L sub-LSPs compared to the second incoming interface that is causing the re-merge condition.

The RSVP-TE Notify messages do not include S2L_SUB_LSP objects and cannot be used to send errors for a subset of the S2L sub-LSPs in a Path message. For that reason, the error generating node SHOULD use a PathErr message rather than a Notify message to communicate the error. In the case of a re-merge error, the node SHOULD use the error code "Routing Problem" and the error value "ERO resulted in re-merge" as specified in [RFC4875].

A border node receiving a PathErr message for a set of S2L sub-LSPs MAY hold the message and attempt to signal an alternate path that can avoid re-merge through its domain for those S2L sub-LSPs that pass through it. However, in the case of a re-merge error for which some of the re-merging S2L sub-LSPs do not pass through the border node, it SHOULD propagate the PathErr upstream towards the ingress node. If the subsequent attempt by the border node is successful, the border node discards the held PathErr and follows the crankback roles of [RFC4920] and [RFC5151]. If repeated subsequent attempts by the border node are unsuccessful, the border node MUST send the held PathErr upstream towards the ingress node.

If the ingress node receives a PathErr message with error code "Routing Problem" and error value "ERO resulted in re-merge", then it SHOULD attempt to signal an alternate path through a different domain or through a different border node for the affected S2L sub-LSPs. The ingress node MAY use the error node information from the PathErr for this purpose.

However, it may be that the ingress node or an ingress border node does not have sufficient topology information to compute an Explicit Route that is guaranteed to avoid the re-merge link or node. In this case, Route Exclusions [RFC4874] may be particularly helpful. To achieve this, [RFC4874] allows the re-merge information to be presented as route exclusions to force avoidance of the re-merge link or node.

As discussed in [RFC4920] section 3.3, border node MAY keep the history of PathErrs. In case of P2MP LSPs, ingress node and border nodes may keep re-merge PathErrs in history table until S2L sub-LSPs have been successfully established or until local timer expires.

4. Data Plane Solution For Re-merge Handling

As mentioned in [RFC4875], a node may accept the re-merging S2Ls but only send the data from one of these interfaces to its outgoing interfaces. That is, the node MUST drop data from all but one incoming interface causing the re-merge. This ensures that duplicate data is not sent on any outgoing interface. Note that data plane may be either programmed to drop the incoming traffic for the S2L sub-LSP or not programmed at all.

It is desirable to avoid the persistent re-merge condition associated with data plane based solution in the network in order to optimize bandwidth resources in the network.

The following sections define the RSVP-TE signaling extensions for "P2MP- TE Re-merge Recording Request" and "P2MP-TE Re-merge Present" messages.

4.1. P2MP-TE Re-merge Recording Request Flag

In order to indicate to traversed nodes that P2MP-TE re-merge recording is desired, a new flag in the Attribute Flags TLV of the LSP_ATTRIBUTES object defined in [RFC5420] is defined as follows:

Bit Number (to be assigned by IANA): P2MP-TE Re-merge Recording Request flag

The "P2MP-TE Re-merge Recording Request" flag is meaningful in a Path message and is inserted by the ingress node or a border node in the LSP_ATTRIBUTES object.

If the "P2MP-TE Re-merge Recording Request" Flag is set, it implies that the "P2MP-TE Re-merge Present" flag defined in the next section MUST be used to indicate to the ingress and ingress border nodes of the transit domains that a re-merge condition is present for this S2L sub-LSP but accepted, and that incoming traffic is being dropped for this S2L sub-LSP.

The rules of the processing of the Attribute Flags TLV of the LSP_ATTRIBUTES object follow [RFC5420].

4.2. P2MP-TE Re-merge Present Flag

The "P2MP-TE Re-merge Present" Flag is the counter part of the "P2MP-TE Re-merge Recording Request" flag defined above. Specifically, RSVP signaling extension is defined to indicate to the

upstream node of the re-merge condition and that incoming traffic is being dropped for the given S2L.

When a node accepts a re-merge condition by dropping traffic from an incoming interface for an S2L due to the re-merge condition, and if it understands the "P2MP-TE Re-merge Recording Request" in the Attribute Flags TLV of the LSP_ATTRIBUTES object of the Path message, the node MUST set the newly defined "P2MP-TE Re-merge Present" flag in the RRO Attributes sub-object defined in [RFC5420] in RRO.

The following new flag for RRO Attributes Sub-object is defined as follows:

Bit Number (same as bit number assigned for "P2MP-TE Re-merge Recording Request" flag): P2MP-TE Re-merge Present flag

The "P2MP-TE Re-merge Present" flag indicates that the S2L is causing a re-merge. The re-merge has been accepted but the incoming traffic on this S2L is dropped by the reporting node.

The rules of the processing of the RRO Attribute Sub-object in the Resv message follow [RFC5420].

4.3. Signaling Procedure

When a node that does not support data plane based re-merge handling receives an S2L sub-LSP Path message with LSP Attributes sub-object that has "P2MP-TE Re-merge Recording Request" Flag set, and if the S2L is causing a re-merge condition, the node MUST reject the S2L sub-LSP Path message and send the PathErr with the error code "Routing Problem" and the error value "ERO resulted in re-merge" as specified in [RFC4875]. If a node is capable of data plane based re-merge handling but operator may have disabled it via a configuration, the the node MUST also reject the re-merge and send this PathErr.

When a Path message is received at a transit node for an S2L sub-LSP and "P2MP-TE Re-merge Recording Request" Flag is set in the LSP Attributes sub-object, the node MAY decide to accept the re-merge S2L sub-LSP based on the local policy and node capability. In this case, before the Resv message is sent to the upstream node for this S2L sub-LSP, the node MUST add the RRO Attributes sub-object in the Resv RRO if not already present and set the "P2MP-TE Re-merge Present" Flag if traffic from the incoming interface of this S2L sub-LSP will be dropped. This same incoming interface can still be used for a different S2L sub-LSP in the P2MP LSP to forward traffic and "P2MP-TE Re-merge Present" flag will not be set for that S2L sub-LSP. Note

that rules for adding or modifying the other RRO sub-objects do not change due to this flag.

When a transit node receives a Resv message for an S2L that is causing a re-merge condition, the node **MUST** set the "P2MP-TE Re-merge Present" flag in the RRO Attributes sub-object in the Resv message if it decides to drop the incoming traffic of this S2L. The "P2MP-TE Re-merge Present" flag in RRO Attribute sub-object is not set for the S2L(s) whose incoming interface is selected to receive and forward the traffic.

An ingress node **MAY** immediately start sending traffic on all S2Ls in up state even when re-merge conditions are present on some S2Ls of the P2MP LSP.

The proposed signaling extensions allow an ingress node and an ingress border node to have a complete view of the re-merge conditions on the entire S2L path and on all S2Ls of the P2MP tree. The ingress or ingress border node in this case can take appropriate actions to resolve the re-merge conditions and optimize network bandwidth resources usage. This can be achieved by computing and selecting alternate path(s) for the S2L(s) bypassing the re-merge node(s).

The proposed signaling extensions are equally applicable to single domain scenarios.

A node where re-merge is present, may decide to select a different incoming interface to forward traffic from in the future. In that case, a Resv change message with updated "P2MP-TE Re-merge Present" flag in the RRO is sent upstream for all effected S2Ls. For the new set of S2L sub-LSPs whose traffic from the incoming interface is dropped, "P2MP-TE Re-merge Present" flag will be set.

A border node due to local policy **MAY** remove the record route object from the Resv message of the S2L sub-LSP and propagate Resv message towards the ingress node. When such a policy is provisioned, the border node may attempt to correct the re-merge condition in its domain. If the border node is not able to resolve the re-merge condition, the border node **SHOULD** send the PathErr with the error code "Routing Problem" and the error value "ERO resulted in re-merge" as specified in [RFC4875].

5. Intra-domain P2MP-TE LSP Re-merge Handling

Re-merges between S2Ls in a single domain can occur due to provisioning errors or path computation errors in the environment

where IGP-TE or PCE is used. Re-merges can also happen in the environment where static routing or static path selection policy is applied at ingress (e.g., CSPF calculation is disabled due to some operational reasons), regardless of using loose or static hops. In either case, procedures described in this document are equally applicable to the intra-domain (i.e. single domain) P2MP-TE LSPs.

6. Reoptimization Handling

6.1. P2MP-TE Tree Re-evaluation Request Flag

In order to query border nodes to check if a preferable P2MP tree exists, a new flag is defined in Attributes Flags TLV of the LSP_ATTRIBUTES object [RFC5420] as follows:

Bit Number (to be assigned by IANA): P2MP-TE Tree Re-evaluation Request flag

The "P2MP-TE Tree Re-evaluation Request" flag is meaningful in a Path message of an S2L sub-LSP and is inserted by the ingress node.

6.2. Preferable P2MP-TE Tree Exists Flag

In order to indicate to an ingress node that a preferable P2MP-TE tree is available, following new sub-code for PathErr code 25 (notify error) is defined:

Sub-code (to be assigned by IANA): Preferable P2MP-TE Tree Exists flag

When a preferable P2MP-TE tree is found, the border node MUST send "Preferable P2MP-TE Tree Exists" to the ingress node in order to reoptimize the entire P2MP LSP.

6.3. Signaling Procedure

Using signaling procedure defined in [RFC4736], an ingress node MUST initiate "path re-evaluation request" query to reoptimize a destination in a P2MP LSP. Note that this message MUST be used to reoptimize a single or a sub-set of the destinations in a P2MP LSP. Ingress node MUST send this query in a Path message for each destination it is reoptimizing.

When a Path message for a destination in a P2MP LSP with "path re-evaluation request" flag [RFC4736] is received at the border node,

it MUST re-compute the loose-hop ERO to see if a preferable path exists for that destination. A border node MUST send PathErr code 25 (notify error defined in [RFC3209]) with "preferable path exists" sub-code to indicate that a preferable path exists for the requested destination AND border node is capable of per destination reoptimization. A border node MUST terminate the path query. Alternatively, a border node not capable of per destination reoptimization MAY respond with "Preferable P2MP-TE Tree Exists" PathErr by checking for a preferable P2MP tree instead of a preferable single destination.

It is often desired to reoptimize the entire P2MP LSP. In order to query border nodes to check if a preferable P2MP tree exists, an ingress node MUST send a Path message with "P2MP-TE Tree Re-evaluation Request" defined in this document. An ingress node MAY send this message for all destinations in a P2MP LSP or a sub-set of the destinations.

A border node receiving the "P2MP-TE Tree Re-evaluation Request" MUST check for a preferable P2MP LSP for the destinations it is loosely routing by loose-hop ERO expansions. The border node if a preferable P2MP-TE tree is found, MUST reply with "Preferable P2MP-TE Tree Exists" sub-code defined in this document with PathErr 25 (notify error defined in [RFC3209]) and terminate the path query.

Note that a border node MAY send "Preferable P2MP-TE Tree Exists" with PathErr code 25 to indicate the ingress node in order to reoptimize the entire P2MP LSP message unsolicited or in a response to "path re-evaluation query" for a destination or in a response to "P2MP-TE Tree Re-evaluation Request" message.

If an ingress node initiated a "path re-evaluation request" query for a single destination for per S2L sub-LSP reoptimization and receives "Preferable P2MP-TE Tree Exists" PathErr, the ingress node MAY cancel the per S2L reoptimization and initiate P2MP-TE tree reoptimization. This may happen in case when a border node is not capable of per destination reoptimization.

Note that even if per destination reoptimization, not whole P2MP LSP Tree reoptimization, is sufficient, ingress node often needs to re-signal whole P2MP LSP tree to complete route optimization for that destination. In this case, make-before-break reoptimization scheme is used (see [RFC4875] Section 14.1), and all S2L sub-LSPs are re-signaled with a different LSP-ID. That is, the procedure of signaling a re-optimization by an ingress node is separate from the matter if PathErr reply was "Preferable Path Exists" or "Preferable P2MP-TE Tree Exists".

7. Compatibility

The LSP_ATTRIBUTES TLV and RRO Attributes sub-object have been defined [RFC5420] with class numbers in the form 1lbbbbbb, which ensures compatibility with non-supporting nodes. Per [RFC2205], nodes not supporting this extension will ignore the TLV, sub-object and the new flags defined in this document but forward it, unexamined and unmodified, in all messages resulting from this message.

8. Security Considerations

This document does not introduce any additional security issues above those identified in [RFC3209], [RFC4875], [RFC5151], [RFC4920] and [RFC5920].

9. IANA Considerations

The following new flag is defined for the Attributes Flags TLV in the LSP_ATTRIBUTES object [RFC5420]. The numeric values are to be assigned by IANA.

- o P2MP-TE Re-merge Recording Request Flag:

- Bit Number: To be assigned by IANA.
- Attribute flag carried in Path message: Yes
- Attribute flag carried in Resv message: No

The following new flag is defined for the RRO Attributes sub-object in the RECORD_ROUTE object [RFC5420]. The numeric values are to be assigned by IANA.

- o P2MP-TE Re-merge Present Flag:

- Bit Number: To be assigned by IANA.
- Attribute flag carried in Path message: No
- Attribute flag carried in RRO Attributes sub-object in RRO of the Resv message: Yes

The following new flag is defined for the Attributes Flags TLV in the LSP_ATTRIBUTES object [RFC5420]. The numeric values are to be assigned by IANA.

- o P2MP-TE Tree Re-evaluation Request Flag:
 - Bit Number: To be assigned by IANA.
 - Attribute flag carried in Path message: Yes
 - Attribute flag carried in Resv message: No

As defined in [RFC3209], the Error Code 25 in the ERROR_SPEC object corresponds to a Notify Error PathErr. This document adds a new sub-code as follows for this PathErr:

- o Preferable P2MP-TE Tree Exists sub-code:
 - Sub-code for Notify PathErr code 25. To be assigned by IANA.

10. Acknowledgments

The authors would like to thank N. Neate for his contributions on the draft.

11. References

11.1. Normative References

- [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.
- [RFC5151] Farrel, A., Ayyangar, A., and JP. Vasseur, "Inter-Domain MPLS and GMPLS Traffic Engineering -- Resource Reservation Protocol-Traffic Engineering (RSVP-TE) Extensions", RFC 5151, February 2008.
- [RFC4920] Farrel, A., Satyanarayana, A., Iwata, A., Fujita, N., and G. Ash, "Crankback Signaling Extensions for MPLS and GMPLS RSVP-TE", RFC 4920, July 2007.
- [RFC5920] L. Fang, Ed., "Security Framework for MPLS and GMPLS Networks", RFC 5920, July 2010.
- [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.
- [RFC4736] Vasseur, JP., Ikejiri, Y. and Zhang, R, "Reoptimization of Multiprotocol Label Switching (MPLS) Traffic Engineering (TE) Loosely Routed Label Switched Path (LSP)", RFC 4736, November 2006.
- [RFC5420] Farrel, A., Papadimitriou, D., Vasseur, JP., and A. Ayyangar, "Encoding of Attributes for MPLS LSP Establishment Using Resource Reservation Protocol Traffic Engineering (RSVP-TE)", RFC 5420, February 2009.

11.2. Informative References

- [RFC4726] Farrel, A., Vasseur, J., and A. Ayyangar, "A Framework for Inter-Domain Multiprotocol Label Switching Traffic Engineering", RFC 4726, November 2006.

Author's Addresses

Zafar Ali
Cisco Systems

Email: zali@cisco.com

Rakesh Gandhi
Cisco Systems

Email: rgandhi@cisco.com

Tarek Saad
Cisco Systems

Email: tsaad@cisco.com

Robert H. Venator
Defense Information Systems Agency

Email: robert.h.venator.civ@mail.mil

Yuji Kamite
NTT Communications Corporation

Email: y.kamite@ntt.com

MPLS
Internet-Draft
Intended status: Informational
Expires: August 23, 2013

C. Villamizar, Ed.
Outer Cape Cod Network
Consulting
February 19, 2013

Use of Multipath with MPLS-TP and MPLS
draft-ietf-mpls-multipath-use-00

Abstract

Many MPLS implementations have supported multipath techniques and many MPLS deployments have used multipath techniques, particularly in very high bandwidth applications, such as provider IP/MPLS core networks. MPLS-TP has strongly discouraged the use of multipath techniques. Some degradation of MPLS-TP OAM performance cannot be avoided when operating over many types of multipath implementations.

Using MPLS Entropy label, MPLS LSPs can be carried over multipath links while also providing a fully MPLS-TP compliant server layer for MPLS-TP LSPs. This document describes the means of supporting MPLS as a server layer for MPLS-TP. The use of MPLS-TP LSPs as a server layer for MPLS LSPs is also discussed.

Status of this Memo

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

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

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

This Internet-Draft will expire on August 23, 2013.

Copyright Notice

Copyright (c) 2013 IETF Trust and the persons identified as the document authors. All rights reserved.

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

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

Table of Contents

| | |
|--|----|
| 1. Introduction | 3 |
| 2. Definitions | 3 |
| 3. MPLS as a Server Layer for MPLS-TP | 5 |
| 3.1. MPLS-TP Forwarding and Server Layer Requirements | 6 |
| 3.2. Methods of Supporting MPLS-TP client LSPs over MPLS | 7 |
| 4. MPLS-TP as a Server Layer for MPLS | 10 |
| 5. Acknowledgements | 11 |
| 6. Implementation Status | 11 |
| 7. IANA Considerations | 11 |
| 8. Security Considerations | 12 |
| 9. References | 12 |
| 9.1. Normative References | 12 |
| 9.2. Informative References | 12 |
| Author's Address | 13 |

1. Introduction

Today the requirement to handle large aggregations of traffic, can be handled by a number of techniques which we will collectively call multipath. Multipath applied to parallel links between the same set of nodes includes Ethernet Link Aggregation [IEEE-802.1AX], link bundling [RFC4201], or other aggregation techniques some of which may be vendor specific. Multipath applied to diverse paths rather than parallel links includes Equal Cost MultiPath (ECMP) as applied to OSPF, ISIS, or BGP, and equal cost LSPs. Some vendors support load split across equal cost MPLS LSPs where the load is split proportionally to the reserved bandwidth of the set of LSPs.

RFC 5654 requirement 33 requires the capability to carry a client MPLS-TP or MPLS layer over a server MPLS-TP or MPLS layer [RFC5654]. This is possible in all cases with one exception. When an MPLS LSP exceeds the capacity of any single component link it may be carried by a network using multipath techniques, but may not be carried by a single MPLS-TP LSP due to the inherent MPLS-TP capacity limitation imposed by MPLS-TP OAM fate sharing constraints and MPLS-TP LM OAM packet ordering constraints (see Section 3.1).

The term composite link is more general than terms such as link aggregation (which is specific to Ethernet) or ECMP (which implies equal cost paths within a routing protocol). The use of the term composite link here is consistent with the broad definition in [ITU-T.G.800]. Multipath is very similar to composite link as defined by ITU, but specifically excludes inverse multiplexing.

2. Definitions

Multipath

The term multipath includes all techniques in which

1. Traffic can take more than one path from one node to a destination.
2. Individual packets take one path only. Packets are not subdivided and reassembled at the receiving end.
3. Packets are not resequenced at the receiving end.
4. The paths may be:
 - a. parallel links between two nodes, or

- b. may be specific paths across a network to a destination node, or
- c. may be links or paths to an intermediate node used to reach a common destination.

Link Bundle

Link bundling is a multipath technique specific to MPLS [RFC4201]. Link bundling supports two modes of operations. Either an LSP can be placed on one component link of a link bundle, or an LSP can be load split across all members of the bundle. There is no signaling defined which allows a per LSP preference regarding load split, therefore whether to load split is generally configured per bundle and applied to all LSPs across the bundle.

Link Aggregation

The term "link aggregation" generally refers to Ethernet Link Aggregation [IEEE-802.1AX] as defined by the IEEE. Ethernet Link Aggregation defines a Link Aggregation Control Protocol (LACP) which coordinates inclusion of LAG members in the LAG.

Link Aggregation Group (LAG)

A group of physical Ethernet interfaces that are treated as a logical link when using Ethernet Link Aggregation is referred to as a Link Aggregation Group (LAG).

Equal Cost Multipath (ECMP)

Equal Cost Multipath (ECMP) is a specific form of multipath in which the costs of the links or paths must be equal in a given routing protocol. The load may be split equally across all available links (or available paths), or the load may be split proportionally to the capacity of each link (or path).

Loop Free Alternate Paths

"Loop-free alternate paths" (LFA) are defined in RFC 5714, Section 5.2 [RFC5714] as follows. "Such a path exists when a direct neighbor of the router adjacent to the failure has a path to the destination that can be guaranteed not to traverse the failure." Further detail can be found in [RFC5286]. LFA as defined for IPFRR can be used to load balance by relaxing the equal cost criteria of ECMP, though IPFRR defined LFA for use in selecting protection paths. When used with IP, proportional split is generally not used. LFA use in load balancing is implemented by some vendors though it may be rare or non-existent in deployments.

Composite Link

The term Composite Link had been a registered trademark of Avici Systems, but was abandoned in 2007. The term composite link is now defined by the ITU in [ITU-T.G.800]. The ITU definition includes multipath as defined here, plus inverse multiplexing which is explicitly excluded from the definition of multipath.

Inverse Multiplexing

Inverse multiplexing either transmits whole packets and resequences the packets at the receiving end or subdivides packets and reassembles the packets at the receiving end. Inverse multiplexing requires that all packets be handled by a common egress packet processing element and is therefore not useful for very high bandwidth applications.

Component Link

The ITU definition of composite link in [ITU-T.G.800] and the IETF definition of link bundling in [RFC4201] both refer to an individual link in the composite link or link bundle as a component link. The term component link is applicable to all multipath.

LAG Member

Ethernet Link Aggregation as defined in [IEEE-802.1AX] refers to an individual link in a LAG as a LAG member. A LAG member is a component link. An Ethernet LAG is a composite link. IEEE does not use the terms composite link or component link.

load split

Load split, load balance, or load distribution refers to subdividing traffic over a set of component links such that load is fairly evenly distributed over the set of component links and certain packet ordering requirements are met. Some existing techniques better achieve these objectives than others.

A small set of requirements are discussed. These requirements make use of keywords such as MUST and SHOULD as described in [RFC2119].

3. MPLS as a Server Layer for MPLS-TP

An MPLS LSP may be used as a server layer for MPLS-TP LSPs as long as all MPLS-TP requirements are met. Section 3.1 reviews the basis for requirements of a server layer that supports MPLS-TP as a client layer. Key requirements include OAM "fate-sharing" the requirement that packets within an MPLS-TP LSP are not reordered, including both payload and OAM packets. Section 3.2 discusses implied requirements where MPLS is the server layer for MPLS-TP

client LSPs, and describes a set of solutions using existing MPLS mechanisms.

3.1. MPLS-TP Forwarding and Server Layer Requirements

[RFC5960] defines the data plane requirements for MPLS-TP. Two very relevant paragraphs in "Section 3.1.1 LSP Packet Encapsulation and Forwarding" are the following.

RFC5960, Section 3.1.1, Paragraph 3

Except for transient packet reordering that may occur, for example, during fault conditions, packets are delivered in order on L-LSPs, and on E-LSPs within a specific ordered aggregate.

RFC5960, Section 3.1.1, Paragraph 6

Equal-Cost Multi-Path (ECMP) load-balancing MUST NOT be performed on an MPLS-TP LSP. MPLS-TP LSPs as defined in this document MAY operate over a server layer that supports load-balancing, but this load-balancing MUST operate in such a manner that it is transparent to MPLS-TP. This does not preclude the future definition of new MPLS-TP LSP types that have different requirements regarding the use of ECMP in the server layer.

[RFC5960] paragraph 3 requires that packets within a specific ordered aggregate be delivered in order. This same requirement is already specified by Differentiated Services [RFC2475]. [RFC5960] paragraph 6 explicitly allows a server layer to use ECMP provided that it is transparent to the MPLS-TP client layer.

[RFC6371] adds a requirement for data traffic and OAM traffic "fate-sharing". The following paragraph in "Section 1 Introduction" summarizes this requirement.

RFC6371, Section 1, Paragraph 7

OAM packets that instrument a particular direction of a transport path are subject to the same forwarding treatment (i.e., fate-share) as the user data packets and in some cases, where Explicitly TC-encoded-PSC LSPs (E-LSPs) are employed, may be required to have common per-hop behavior (PHB) Scheduling Class (PSC) End-to-End (E2E) with the class of traffic monitored. In case of Label-Only-Inferred-PSC LSP (L-LSP), only one class of traffic needs to be monitored, and therefore the OAM packets have common PSC with the monitored traffic class.

[RFC6371] does not prohibit multilink techniques in "Section 4.6 Fate-Sharing Considerations for Multilink", where multilink is defined as Ethernet Link Aggregation and the use of Link Bundling for MPLS, but does declare that such a network would be only partially

MPLS-TP compliant. The characteristic that is to be avoided is contained in the following sentence in this section.

RFC6371, Section 4.6, Paragraph 1, last sentence

These techniques frequently share the characteristic that an LSP may be spread over a set of component links and therefore be reordered, but no flow within the LSP is reordered (except when very infrequent and minimally disruptive load rebalancing occurs).

A declaration that implies that Link Bundling for MPLS yields a partially MPLS-TP compliant network, is perhaps overstated since only the Link Bundling all-ones component link has this characteristic.

[RFC6374] defines a direct Loss Measurement (LM) where LM OAM packets cannot be reordered with respect to payload packets. This will require that payload packets themselves not be reordered. The following paragraph in "Section 2.9.4 Equal Cost Multipath" gives the reason for this restriction.

RFC6374, Section 2.9.4, Paragraph 2

The effects of ECMP on loss measurement will depend on the LM mode. In the case of direct LM, the measurement will account for any packets lost between the sender and the receiver, regardless of how many paths exist between them. However, the presence of ECMP increases the likelihood of misordering both of LM messages relative to data packets and of the LM messages themselves. Such misorderings tend to create unmeasurable intervals and thus degrade the accuracy of loss measurement. The effects of ECMP are similar for inferred LM, with the additional caveat that, unless the test packets are specially constructed so as to probe all available paths, the loss characteristics of one or more of the alternate paths cannot be accounted for.

3.2. Methods of Supporting MPLS-TP client LSPs over MPLS

Supporting MPLS-TP LSPs over a fully MPLS-TP conformant MPLS LSP server layer where the MPLS LSPs are making use of multipath, requires special treatment of the MPLS-TP LSPs such that those LSPs meet MPLS-TP forwarding requirements (see Section 3.1). This implies the following brief set of requirements.

MP#1 It MUST be possible for a midpoint MPLS-TP LSR which is serving as ingress to a server layer MPLS LSP to identify MPLS-TP LSPs, so that MPLS-TP forwarding requirements can be applied, or to otherwise accommodate the MPLS-TP forwarding requirements.

- MP#2 It SHOULD be possible to completely exclude MPLS-TP LSPs from the multipath hash and load split. If the selected component link no longer meets requirements, an LSP is considered down which may trigger protection and/or may require that the ingress LSR select a new path and signal a new LSP.
- MP#3 It SHOULD be possible to insure that MPLS-TP LSPs will not be moved to another component link as a result of a composite link load rebalancing operation. If the selected component link no longer meets requirements, another component link may be selected, however a change in path should not occur solely for load balancing.
- MP#4 Where an RSVP-TE control plane is used, it MUST be possible for an ingress LSR which is setting up an MPLS-TP or an MPLS LSP to determine at path selection time whether a link or Forwarding Adjacency (FA, see [RFC4206]) within the topology can support the MPLS-TP requirements of the LSP.

The reason for requirement MP#1 may not be obvious. A MPLS-TP LSP may be aggregated along with other client LSP by a midpoint LSR into a very large MPLS server layer LSP, as would be the case in a core node to core node MPLS LSP between major cities. In this case the ingress of the MPLS LSP cannot through any existing signaling mechanism determine which client LSP contained within it as MPLS-TP or not MPLS-TP. Multipath load splitting can be avoided for MPLS-TP LSP if at the MPLS server layer LSP ingress LSR an Entropy Label Indicator (ELI) and Entropy Label (EL) are added to the label stack [RFC6790]. For those client LSP that are MPLS-TP LSP, a single EL value must be chosen. For those client LSP that are MPLS LSP, per packet entropy below the top label must, for practical reasons, be used to determine the entropy label value. Requirement MP#1 simply states that there must be a means to make this decision.

There is currently no signaling mechanism defined to support requirement MP#1, though that does not preclude a new extension being defined later. In the absense of a signaling extension, MPLS-TP can be identified through some form of configuration, such as configuration which provides an MPLS-TP compatible server layer to all LSP arriving on a specific interface or originating from a specific set of ingress LSR.

Alternately, the need for requirement MP#1 can be eliminated if every MPLS-TP LSP can be created by the MPLS-TP ingress makes use of an Entropy Label Indicator (ELI) and Entropy Label (EL) below the MPLS-TP label [RFC6790]. This would require that all MPLS-TP LSR in a deployment support Entropy Label, which may render it impractical in many deployments.

Some hardware which exists today can support requirement MP#2. Signaling in the absence of MPLS Entropy Label can make use of link bundling with the path pinned to a specific component for MPLS-TP LSP and link bundling using the all-ones component for MPLS LSP. This prevents MPLS-TP LSP from being carried within MPLS LSP but does allow the co-existence of MPLS-TP and very large MPLS LSP.

MPLS-TP LSPs can be carried as client LSPs within an MPLS server LSP if an Entropy Label Indicator (ELI) and Entropy Label (EL) is added after the server layer LSP label(s) in the label stack, just above the MPLS-TP LSP label entry [RFC6790]. The value of EL can be randomly selected at the client MPLS-TP LSP setup time and the same EL value used for all packets of that MPLS-TP LSP. This allows MPLS-TP LSP to be carried as client LSP within MPLS LSP and satisfies MPLS-TP forwarding requirements but requires that MPLS LSR be able to identify MPLS-TP LSP (requirement MP#1).

MPLS-TP traffic can be protected from a degraded performance due to an imperfect load split if the MPLS-TP traffic is given queuing priority (using strict priority and policing or shaping at ingress or locally or weighted queuing locally). This can be accomplished using the Traffic Class field and Diffserv treatment of traffic [RFC5462][RFC2475]. In the event of congestion due to load imbalance, other traffic will suffer as long as there is a minority of MPLS-TP traffic.

If MPLS-TP LSP are carried within MPLS LSP and ELI and EL are used, requirement MP#3 is satisfied only for uncongested links where load balancing is not required, or if MPLS-TP LSP use TC and Diffserv and the load rebalancing implementation rebalances only the less preferred traffic. Load rebalance is generally needed only when congestion occurs, therefore restricting MPLS-TP to be carried only over MPLS LSP that are known to traverse only links which are expected to be uncongested can satisfy requirement MP#3.

An MPLS-TP LSP can be pinned to a Link Bundle component link if the behavior of requirement MP#2 is preferred. An MPLS-TP LSP can be assigned to a Link Bundle but not pinned if the behavior of requirement MP#3 is preferred. In both of these cases, the MPLS-TP LSP must be the top level LSP, except as noted above.

If MPLS-TP LSP can be moved among component links, then the Link Bundle all-ones component link can be used or server layer MPLS LSPs can be used with no restrictions on the server layer MPLS use of multipath except that Entropy Label must be supported along the entire path. An Entropy Label must be used to insure that all of the MPLS-TP payload and OAM traffic are carried on the same component, except during very infrequent transitions due to load balancing.

An MPLS-TP LSP may not traverse multipath links on the path where MPLS-TP forwarding requirements cannot be met. Such links include any using pre-RFC6790 Ethernet Link Aggregation, pre-RFC6790 Link Bundling using the all-ones component link, or other form of multipath not supporting termination of the entropy search at the EL label as called for in [RFC6790]. An MPLS-TP LSP must not traverse a server layer MPLS LSP which traverses any form of multipath not supporting termination of the entropy search at the EL label. For this to occur, the MPLS-TP ingress LSR must be aware of these links. This is the reason for requirement MP#4.

Requirement MP#4 can be supported using administrative attributes. Administrative attributes are defined in [RFC3209]. Some configuration is required to support this.

4. MPLS-TP as a Server Layer for MPLS

Carrying MPLS LSP which are larger than a component link over a MPLS-TP server layer requires that the large MPLS client layer LSP be accommodated by multiple MPLS-TP server layer LSPs. MPLS multipath can be used in the client layer MPLS.

Creating multiple MPLS-TP server layer LSP places a greater Incoming Label Map (ILM) scaling burden on the LSR. High bandwidth MPLS cores with a smaller amount of nodes have the greatest tendency to require LSP in excess of component links, therefore the reduction in number of nodes offsets the impact of increasing the number of server layer LSP in parallel. Today, only in cases where deployed LSR ILM are small would this be an issue.

The most significant disadvantage of MPLS-TP as a Server Layer for MPLS is that the use MPLS-TP server layer LSP reduces the efficiency of carrying the MPLS client layer. The service which provides by far the largest offered load in provider networks is Internet, for which the LSP capacity reservations are predictions of expected load. Many of these MPLS LSP may be smaller than component link capacity. Using MPLS-TP as a server layer results in bin packing problems for these smaller LSP. For those LSP that are larger than component link capacity, their capacity are not increments of convenient capacity increments such as 10Gb/s. Using MPLS-TP as an underlying server layer greatly reduces the ability of the client layer MPLS LSP to share capacity. For example, when one MPLS LSP is underutilizing its predicted capacity, the fixed allocation of MPLS-TP to component links may not allow another LSP to exceed its predicted capacity. Using MPLS-TP as a server layer may result in less efficient use of resources and may result in a less cost effective network.

No additional requirements beyond MPLS-TP as it is now currently defined are required to support MPLS-TP as a Server Layer for MPLS. It is therefore viable but has some undesirable characteristics discussed above.

5. Acknowledgements

Carlos Pignataro, Dave Allan, and Mach Chen provided valuable comments and suggestions. Carlos suggested that MPLS-TP requirements in RFC 5960 be explicitly referenced or quoted. An email conversation with Dave led to the inclusion of references and quotes from RFC 6371 and RFC 6374. Mach made suggestions to improve clarity of the document.

6. Implementation Status

Note: this section is temporary and supports the experiment called for in draft-sheffer-running-code.

This is an informational document which describes usage of MPLS and MPLS-TP. No new protocol extensions or forwarding behavior are specified. Ethernet Link Aggregation and MPLS Link Bundling are widely implemented and deployed.

Entropy Label is not yet widely implemented and deployed, but both implementation and deployment are expected soon. At least a few existing high end commodity packet processing chips are capable of supporting Entropy Label. It would be helpful if a few LSR suppliers would state their intentions to support RFC 6790 on the mpls mailing list.

Dynamic multipath (multipath load split adjustment in response to observed load) is referred to but not a requirement of the usage recommendations made in this document. Dynamic multipath has been implemented and deployed, however (afaik) the only core LSR vendor supporting dynamic multipath is no longer in the router business (Avici Systems). At least a few existing high end commodity packet processing chips are capable of supporting dynamic multipath.

7. IANA Considerations

This memo includes no request to IANA.

8. Security Considerations

This document specifies requirements with discussion of framework for solutions using existing MPLS and MPLS-TP mechanisms. The requirements and framework are related to the coexistence of MPLS/GMPLS (without MPLS-TP) when used over a packet network, MPLS-TP, and multipath. The combination of MPLS, MPLS-TP, and multipath does not introduce any new security threats. The security considerations for MPLS/GMPLS and for MPLS-TP are documented in [RFC5920] and [I-D.ietf-mpls-tp-security-framework].

9. References

9.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC5654] Niven-Jenkins, B., Brungard, D., Betts, M., Sprecher, N., and S. Ueno, "Requirements of an MPLS Transport Profile", RFC 5654, September 2009.
- [RFC5960] Frost, D., Bryant, S., and M. Bocci, "MPLS Transport Profile Data Plane Architecture", RFC 5960, August 2010.
- [RFC6371] Busi, I. and D. Allan, "Operations, Administration, and Maintenance Framework for MPLS-Based Transport Networks", RFC 6371, September 2011.
- [RFC6374] Frost, D. and S. Bryant, "Packet Loss and Delay Measurement for MPLS Networks", RFC 6374, September 2011.
- [RFC6790] Kompella, K., Drake, J., Amante, S., Henderickx, W., and L. Yong, "The Use of Entropy Labels in MPLS Forwarding", RFC 6790, November 2012.

9.2. Informative References

- [I-D.ietf-mpls-tp-security-framework]
Fang, L., Niven-Jenkins, B., Mansfield, S., and R. Graveman, "MPLS-TP Security Framework", draft-ietf-mpls-tp-security-framework-05 (work in progress), October 2012.
- [IEEE-802.1AX]
IEEE Standards Association, "IEEE Std 802.1AX-2008 IEEE Standard for Local and Metropolitan Area Networks - Link

Aggregation", 2006, <<http://standards.ieee.org/getieee802/download/802.1AX-2008.pdf>>.

[ITU-T.G.800]

ITU-T, "Unified functional architecture of transport networks", 2007, <<http://www.itu.int/rec/T-REC-G/recommendation.asp?parent=T-REC-G.800>>.

- [RFC2475] Blake, S., Black, D., Carlson, M., Davies, E., Wang, Z., and W. Weiss, "An Architecture for Differentiated Services", RFC 2475, December 1998.
- [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.
- [RFC4201] Kompella, K., Rekhter, Y., and L. Berger, "Link Bundling in MPLS Traffic Engineering (TE)", RFC 4201, October 2005.
- [RFC4206] Kompella, K. and Y. Rekhter, "Label Switched Paths (LSP) Hierarchy with Generalized Multi-Protocol Label Switching (GMPLS) Traffic Engineering (TE)", RFC 4206, October 2005.
- [RFC5286] Atlas, A. and A. Zinin, "Basic Specification for IP Fast Reroute: Loop-Free Alternates", RFC 5286, September 2008.
- [RFC5462] Andersson, L. and R. Asati, "Multiprotocol Label Switching (MPLS) Label Stack Entry: "EXP" Field Renamed to "Traffic Class" Field", RFC 5462, February 2009.
- [RFC5714] Shand, M. and S. Bryant, "IP Fast Reroute Framework", RFC 5714, January 2010.
- [RFC5920] Fang, L., "Security Framework for MPLS and GMPLS Networks", RFC 5920, July 2010.

Author's Address

Curtis Villamizar (editor)
Outer Cape Cod Network Consulting

Email: curtis@ocnc.com

Network Working Group
Internet-Draft
Updates: 3032 (if approved)
Intended status: Standards Track
Expires: April 18, 2013

K. Kompella
Contrail Systems
L. Andersson
Ericsson
A. Farrel
Juniper Networks
October 15, 2012

Allocating and Retiring MPLS Reserved Labels
draft-kompella-mpls-special-purpose-labels-01

Abstract

Some MPLS labels have been allocated for specific purposes. A block of labels (0-15) has been set aside to this end, and are commonly called "reserved labels". They will be called "special purpose labels" in this document. As there are only 16 of these labels, caution is needed in the allocation of new special purpose labels, yet at the same time allow forward progress when one is called for. This memo defines some procedures to follow in the allocation and retirement of special purpose labels, as well as a method to extend the special purpose label space. Finally, this memo renames the IANA registry for these labels to "Special Purpose MPLS Label Values", and creates a new one called the "Extended Special Purpose MPLS Label Values" registry.

Status of this Memo

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

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

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

This Internet-Draft will expire on April 18, 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. Conventions used | 3 |
| 2. Questions | 4 |
| 3. Answers | 5 |
| 3.1. Extended Special Purpose MPLS Label Values | 5 |
| 3.2. Process for Retiring Special Purpose Labels | 6 |
| 4. IANA Considerations | 7 |
| 5. Security Considerations | 8 |
| 6. References | 9 |
| 6.1. Normative References | 9 |
| 6.2. Informational References | 9 |
| Authors' Addresses | 10 |

1. Introduction

The specification of the Label Stack Encoding for Multi-Protocol Label Switching (MPLS) [RFC3032] defined four special purpose label values (0 to 3), and set aside values 4 through 15 for future use. These labels have special significance in both the control and the data plane. Since then, three further values have been allocated (values 7, 13, and 14 in [I-D.ietf-mpls-entropy-label], [RFC5586] and [RFC3429], respectively), leaving nine unassigned values from the original space of sixteen.

While the allocation of three out of the remaining twelve special purpose label values in the space of about 12 years is not in itself a cause for concern, the scarcity of special purpose labels is. Furthermore, many of the special purpose labels require special processing by forwarding hardware, changes to which are often expensive, and sometimes impossible. Thus, documenting a newly allocated special purpose label value is important.

This memo outlines some of the issues in allocating and retiring special purpose label values, and defines mechanisms to address these. This memo also extends the space of special purpose labels.

1.1. Conventions used

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. Questions

In re-appraising MPLS special purpose labels, the following questions come to mind:

1. What allocation policies should be applied by IANA for the allocation of special purpose labels? Should Early Allocation [RFC4020] be allowed? Should there be labels for Experimental Use or Private Use [RFC5226]?
2. What documentation is required for special purpose labels allocated henceforth?
3. Should a special purpose label ever be retired? What criteria are relevant here? Can a retired special purpose label ever be re-allocated for a different purpose? What procedures and time frames are appropriate?
4. The special purpose label value of 3 (the "Implicit Null Label", [RFC3032]) is only used in signaling, never in the data plane. Could it (and should it) be used in the data plane? If so, how and for what purpose?
5. What is a feasible mechanism to extend the space of special purpose labels, should this become necessary?

3. Answers

This section provides answers to the questions posed in the previous section.

1.

- A. Allocation of special purpose MPLS labels is via "Standards Action".
- B. The IANA registry will be renamed "Special Purpose MPLS Labels".
- C. Early allocation may be allowed on a case-by-case basis.
- D. The current space of 16 special purpose labels is too small for setting aside value for experimental or private use. However, the extended special purpose labels registry created by this document has enough space, and this document defines a range for experimental use.

- 2. A Standards Track RFC must accompany a request for allocation of special purpose labels, as per [RFC5226].
- 3. The retirement of a special purpose MPLS label value must follow a strict and well-documented process. This is necessary since we must avoid orphaning the use of this label value in existing deployments. This process is detailed in Section 3.2.
- 4. The use of the "implicit null label" (label 3) in the data plane may be allowed, subject to approval by the MPLS WG, and an accompanying Standards Track RFC that details the use of the label, and a discussion of possible sources of confusion between signaling and data plane, and mitigation thereof.
- 5. The special purpose label (the "extension" label) is to be set aside for the purpose of extending the space of special purpose labels. Further details are described in Section 3.1.

A further question to be settled in this regard is whether a "regular" special purpose label retains its meaning if it follows the extension label; see Section 3.1.

3.1. Extended Special Purpose MPLS Label Values

An extension label MUST be followed by another label L (and thus MUST have the bottom-of-stack bit clear). L MUST be interpreted as an "extended special purpose label" from a new registry created by this

document (see Section 4). Whether or not L has the bottom-of-stack bit set depends on whether other labels follow L.

IANA is asked to set aside label value 15 as the extension label.

The first 16 values of the extended special purpose label registry are duplicated from the pre-existing special purpose label registry. This includes the previously allocated values (0-3, 7, 13, and 14), the extension label value (15) allocated by this document, and the remaining unallocated values (4-6 and 8-12). Any of these values present as an extended special purpose label MUST be interpreted exactly as it would if it was presented as a special purpose label. In particular, an arbitrary string of consecutive extension labels is legal, and semantically equivalent to a single extension label (note that this string of extension labels MUST be followed by an extended special purpose label that is not the extension label).

3.2. Process for Retiring Special Purpose Labels

- a. A label value that has been assigned from the "Special Purpose MPLS Label Values" may be deprecated by IETF consensus with review by the MPLS working group (or designated experts if the working group or a successor does not exist). An RFC with at least Informational status is required.

The RFC will direct the IANA to mark the label value as "deprecated" in the registry, but will not release it at this stage.

Deprecating means that no further specifications using the deprecated value will be documented.

At the same time this is an indication to vendors not to include deprecated value in new implementations and to operators to avoid including it in new deployments.

- b. 12 months after the RFC deprecating the label value is published, an IETF-wide survey may be conducted to determine if the deprecated label value is still in use. If the survey indicates that the deprecated label value is in use, the survey may be repeated after a further 6 months.
- c. 24 months after the RFC that deprecated the label value was published and if the survey indicates that deprecated label value is not in use, publication may be requested of an IETF Standards Track Internet-Draft that retires the deprecated the label value. This document will request IANA to release the label value for for future use and assignment.

4. IANA Considerations

This document requests IANA to make the following changes and additions to its registration of MPLS Labels.

1. Change the name of the "Multiprotocol Label Switching Architecture (MPLS) Label Values" registry to the "Special Purpose MPLS Label Values".
2. Change the allocations policy for the "Special Purpose MPLS Label Values" registry to Standards Action.
3. Assign label 15 from the "Special Purpose MPLS Label Values" registry, naming it the "extension label", and citing this document as the reference.
4. Create a new registry called the "Extended Special Purpose MPLS Label Values" registry. The ranges and allocation policies for this registry are as follows (using terminology from [RFC5226]). Early allocation following the policy defined in [RFC4020] is allowed only for those values assigned by Standards Action.

| Range | Allocation Policy |
|-------------------|--|
| 0 - 15 | Reserved. Not to be allocated. Meaning is defined by values in the "Special Purpose MPLS Label Values" registry. |
| 16 - 1048559 | Standards Action |
| 1048560 - 1048575 | Experimental |

Table 1

5. Security Considerations

This document does not make a large change to the operation of the MPLS data plane and security considerations are largely unchanged from those specified in the MPLS architecture [RFC3031] and in the MPLS and GMPLS Security Framework [RFC5920].

However, it should be noted that increasing the label stack can cause packet fragmentation and may also make packets unprocessable by some implementations. This document provides a protocol-legal way to arbitrarily increase the label stack and so might provide a way to attack some nodes in a network without violating the protocol rules.

6. References

6.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC3031] Rosen, E., Viswanathan, A., and R. Callon, "Multiprotocol Label Switching Architecture", RFC 3031, January 2001.
- [RFC3032] Rosen, E., Tappan, D., Fedorkow, G., Rekhter, Y., Farinacci, D., Li, T., and A. Conta, "MPLS Label Stack Encoding", RFC 3032, January 2001.
- [RFC4020] Kompella, K. and A. Zinin, "Early IANA Allocation of Standards Track Code Points", BCP 100, RFC 4020, February 2005.
- [RFC5226] Narten, T. and H. Alvestrand, "Guidelines for Writing an IANA Considerations Section in RFCs", BCP 26, RFC 5226, May 2008.
- [RFC5920] Fang, L., "Security Framework for MPLS and GMPLS Networks", RFC 5920, July 2010.

6.2. Informational References

- [RFC3429] Ohta, H., "Assignment of the 'OAM Alert Label' for Multiprotocol Label Switching Architecture (MPLS) Operation and Maintenance (OAM) Functions", RFC 3429, November 2002.
- [RFC5586] Bocci, M., Vigoureux, M., and S. Bryant, "MPLS Generic Associated Channel", RFC 5586, June 2009.
- [I-D.ietf-mpls-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-06 (work in progress), September 2012.

Authors' Addresses

Kireeti Kompella
Contrail Systems
2350 Mission College Blvd.
Santa Clara, CA 95054
US

Email: kireeti.kompella@gmail.com

Loa Andersson
Ericsson

Email: loa@pi.nu

Adrian Farrel
Juniper Networks

Email: adrian@olddog.co.uk

Network Working Group
Internet-Draft
Intended status: Experimental
Expires: August 22, 2013

Z. Li
T. Huang
Huawei Technologies
February 18, 2013

Alternative Constraints for Point-to-Multipoint Traffic-Engineered MPLS
Label Switched Paths(LSPs)
draft-li-mppls-p2mp-te-alt-path-00

Abstract

The document proposes a solution to be able to set up the alternative path for specific leaf nodes of a P2MP TE LSP. Corresponding RSVP-TE protocol extension is also defined. The solution is used to cope with the issue that in some scenarios traffic loss happens even if there exists possible path for the leaf nodes.

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 August 22, 2013.

Copyright Notice

Copyright (c) 2013 IETF Trust and the persons identified as the document authors. All rights reserved.

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

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

Table of Contents

| | |
|--|---|
| 1. Introduction | 3 |
| 2. Terminology | 3 |
| 3. Problem Statement | 3 |
| 4. Mechanisms | 4 |
| 4.1. Path Computation in Root Node | 5 |
| 4.2. Alternative Constraints Propagation | 5 |
| 4.3. Resource Reservation | 6 |
| 5. Protocol Extension | 6 |
| 5.1. Path Message Format | 6 |
| 6. IANA Considerations | 7 |
| 7. Security Considerations | 7 |
| 8. Normative References | 8 |
| Authors' Addresses | 8 |

1. Introduction

[RFC4461] presents a set of requirements for the establishment and maintenance of Point-to-Multipoint (P2MP) Traffic-Engineered (TE) Multiprotocol Label Switching (MPLS) Label Switched Paths (LSPs). [RFC4875] defines extensions to the RSVP-TE protocol for setup of P2MP TE LSPs. P2MP TE LSPs are set up with a series of traffic engineering constraints. These constraints are applied to all S2L sub-LSPs. This may cause the issue that some S2L sub-LSPs can be set up while others can not according to the constraints. There may be worse case that some S2L sub-LSPs can not be restored after link failure according to the constraints. When P2MP TE LSPs are used for specific applications, it will cause continuous traffic loss. This document identifies the applicability issue and proposes the solution and corresponding protocol extension.

2. Terminology

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

3. Problem Statement

The P2MP TE LSP is set up with a series of traffic engineering constraints such as bandwidth, explicit path, affinity property(color), etc. These traffic engineering constraints are applied to path computation for all S2L sub-LSPs. Owing to the network provision some leaves of the P2MP LSP are not reachable according the required constraints (it will be called primary constraints in the following). There may be the worse case that all leaves are reachable at the beginning and they are not reachable when failure happens. In fact these leaves can be reachable if ignore some or all of the primary constraints .

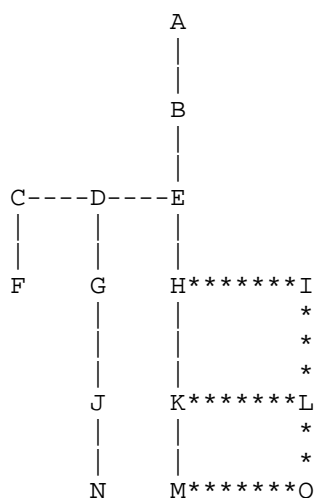


Figure 1. Constraints for P2MP TE LSP

An example for P2MP TE LSP setup is shown in the figure 1. A is the root node and F, N and M are leaf nodes. The link with '|' means the link with red color and the link with '*' means the link with green color. The constraint is that the link with red color should be chosen for the path. For the leaf node M, the path is A->B->E->H->K-M. When link between H and K fails, there is no path with red color can be found from A to M. This will cause the initial available traffic break until the link between H and K restores. The continuous traffic loss can cause bad user experience if the P2MP TE LSP is used for IPTV or other applications. In fact, during the course of failure, there is an alternative path from A to M (A->B->E->H->I->L->K->M) if the link with green color can be chosen.

4. Mechanisms

In order to solve the above applicability issue for P2MP TE LSP, alternative constraints can be specified for the P2MP TE LSP to calculate paths to specific leaf nodes if the path with the primary constraints is not available. The P2MP TE LSP is set up with some S2L sub-LSPs using the primary constraints while the other S2L sub-LSPs using the alternative constraints. The constraints may be used in the downstream nodes, such as ASBR node, and the alternative constraints MUST be propagated to keep the consistence through RSVP-TE protocol extensions.

4.1. Path Computation in Root Node

When alternative constraints is allowed for a specific P2MP TE LSP in the root node, the node MUST try to compute paths for all leaf nodes using the primary constraints. If paths with the primary constraints are available for all leaf nodes, the alternative constraints MUST NOT be used.

When paths with the primary constraints are not available for specific leaf nodes, the alternative constraints SHOULD be used to calculate paths for these leaf nodes. In order to get available paths, the alternative constraints should be looser than the primary constraints. The alternative constraints can be set as zero to simplify the process and the best-effort path as routing is calculated.

When calculate paths with the alternative constraints, the constraints MUST be applied to the whole S2L sub-LSP. That is, it is prohibited that some parts of the S2L sub-LSP satisfies the primary constraints while other parts satisfies the alternative constraints. If the root node can not calculate the whole S2L sub-LSP (abstract node exists in the calculated path), the alternative constraints MUST be used in the downstream nodes path calculation.

The root node will keep trying to re-optimize to a better path to meet the primary constraints, and it is outside the scope of this document.

4.2. Alternative Constraints Propagation

When setup P2MP LSP, the primary constraint is carried according to the RSVP-TE protocol extension which is defined in [RFC4875]. If the paths to specific leaf nodes are computed using alternative constraints, the alternative constraints MUST be carried corresponding to the S2L sub-LSPs to these leaf nodes in the Path message. These alternative constraints corresponding to S2L sub-LSPs are propagated along the paths from the root node to the leaf nodes.

Both the primary and alternative constraints may be propagated in one Path message. a transit node SHOULD choose the correct constraints to calculate the rest path. If there are alternative constraints following the S2L sub-LSPs, it MUST be used when calculating for the S2L sub-LSPs, while the primary constraints MUST be used for the S2L sub-LSPs that is not followed. This will be described in detail in the section 5 of RSVP-TE protocol extensions.

4.3. Resource Reservation

When the Resv message is propagated from the leaf nodes to the root node, the transit node MUST reserve resource according to the traffic parameters specified by the required constraints. However, the common upstream node, such as A, B node in figure 1, may have different traffic parameters required if both the primary and alternative constraints exist, and the primary constraints should be chosen in this case.

5. Protocol Extension

There are two methods for RSVP-TE protocol to carry both the primary and alternative constraints. One is to separate the S2L sub-LSPs with alternative constraints from the S2L sub-LSPs with the primary constraints. The Sub-Group fields imported in [RFC4875] may evade the issue of section 4.2 naturally. It is assumed that the S2L sub-LSPs with the primary constraints and the S2L sub-LSPs with alternative constraints SHOULD not be propagated in a single IP packet. The other method will be described in detail in section 5.1 Path Message Format.

5.1. Path Message Format

This section describes modifications made to the Path message format as specified in [RFC4875]. The Path message is enhanced to signal alternative constraints for specific S2L sub-LSPs.


```

<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>]

```

The following is the format of the S2L sub-LSP descriptor list.

```

<S2L sub-LSP descriptor list> ::= <S2L sub-LSP descriptor>
                                   [ <S2L sub-LSP descriptor list> ]

<S2L sub-LSP descriptor> ::= <S2L_SUB_LSP>
                             [ <P2MP SECONDARY_EXPLICIT_ROUTE> ]
                             [ <P2MP SECONDARY_SESSION_ATTRIBUTE> ]
                             [ <P2MP SECONDARY_SENDER_TSPEC> ]

```

In the modified Path message, S2L_SUB_LSP for specific leaf nodes can carry the alternative constraints besides the explicit route. <P2MP SECONDARY_SESSION_ATTRIBUTE> and <P2MP SECONDARY_SENDER_TSPEC> are added to specify the alternative constraints such as resource affinity, setup and holding priority and traffic parameters. The format of <P2MP SECONDARY_SESSION_ATTRIBUTE> and <P2MP SECONDARY_SENDER_TSPEC> are the same as <SESSION_ATTRIBUTE> defined by [RFC3209] and <SENDER_TSPEC> defined by [RFC2210].

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

TBD.

8. 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.
- [RFC2210] Wroclawski, J., "The Use of RSVP with IETF Integrated Services", RFC 2210, 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.

Authors' Addresses

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

Email: lizhenbin@huawei.com

Tieying Huang
Huawei Technologies
Huawei Bld., No.156 Beiqing Rd.
Beijing 100095
China

Email: huangtieying@huawei.com

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: August 26, 2013

G. Swallow
V. Lim
Cisco Systems
S. Aldrin
Huawei Technologies
February 22, 2013

Proxy MPLS Echo Request
draft-lim-mpls-proxy-lsp-ping-01

Abstract

This document defines a means of remotely initiating Multiprotocol Label Switched Protocol Pings on Label Switched Paths. A proxy ping request is sent to any Label Switching Routers along a Label Switched Path. The primary motivations for this facility are first to limit the number of messages and related processing when using LSP Ping in large Point-to-Multipoint LSPs, and second to enable leaf to leaf/ root tracing.

Status of this Memo

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

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

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

This Internet-Draft will expire on August 26, 2013.

Copyright Notice

Copyright (c) 2013 IETF Trust and the persons identified as the document authors. All rights reserved.

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

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

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

Table of Contents

| | |
|---|----|
| 1. Introduction | 4 |
| 1.1. Requirements Language | 5 |
| 2. Proxy Ping Overview | 5 |
| 3. Proxy MPLS Echo Request / Reply Procedures | 7 |
| 3.1. Procedures for the initiator | 7 |
| 3.2. Procedures for the proxy LSR | 8 |
| 3.2.1. Proxy LSR Handling when it is Egress for FEC | 10 |
| 3.2.2. Downstream Detailed/Downstream Maps in Proxy Reply | 11 |
| 3.2.3. Sending an MPLS proxy ping reply | 11 |
| 3.2.4. Sending the MPLS echo requests | 11 |
| 3.2.4.1. Forming the base MPLS echo request | 11 |
| 3.2.4.2. Per interface sending procedures | 13 |
| 4. Proxy Ping Request / Reply Messages | 13 |
| 4.1. Proxy Ping Request / Reply Message formats | 13 |
| 4.2. Proxy Ping Request Message contents | 14 |
| 4.3. Proxy Ping Reply Message Contents | 15 |
| 5. Object formats | 15 |
| 5.1. Proxy Echo Parameters Object | 16 |
| 5.1.1. Next Hop sub-Object | 19 |
| 5.2. Reply-to Address Object | 20 |
| 5.3. Upstream Neighbor Address Object | 21 |
| 5.4. Downstream Neighbor Address Object | 22 |
| 6. Security Considerations | 23 |
| 7. Acknowledgements | 24 |
| 8. IANA Considerations | 24 |
| 9. References | 25 |
| 9.1. Normative References | 25 |
| 9.2. Informative References | 25 |
| Authors' Addresses | 26 |

1. Introduction

This document is motivated by two broad issues in connection with diagnosing P2MP LSPs. The first is scalability due to the automatic replication of MPLS Echo Request Messages as they proceed down the tree. The second, which is primarily motivated by mLDP, is the ability to trace a sub-LSP from leaf node to root node.

It is anticipated that very large Point-to-Multipoint (P2MP) and Multipoint-to-Multipoint (MP2MP) Label Switched Paths (LSPs) will exist. Further it is anticipated that many of the applications for P2MP/MP2MP tunnels will require OAM that is both rigorous and scalable.

Suppose one wishes to trace a P2MP LSP to localize a fault which is affecting one egress or a set of egresses. Suppose one follows the normal procedure for tracing - namely repeatedly pinging from the root, incrementing the TTL by one after each three or so pings. Such a procedure has the potential for producing a large amount of processing at the P2MP-LSP midpoints and egresses. It also could produce an unwieldy number of replies back to the root.

One alternative would be to begin sending pings from points at or near the affected egress(es) and working backwards toward the root. The TTL could be held constant as say two, limiting the number of responses to the number of next-next-hops of the point where a ping is initiated.

In the case of RSVP-TE, all setup is initiated from the root of the tree. Thus, the root of the tree has knowledge of all the leaf nodes and usually the topology of the entire tree. Thus the above alternative can easily be initiated by the root node.

In mLDP the situation is quite different. Leaf nodes initiate connection to the tree which is granted by the first node that is part of the tree. The root node may only be aware of the immediately adjacent (downstream) nodes of the tree. Initially the leaf node only has knowledge of the node it is immediately adjacent to (upstream) in the tree. However this is sufficient to initiate a trace by applying the above alternative to the last link in the tree. That is, by requesting the upstream node to send an MPLS Echo Request for the FEC of the tree in question on said link. By adding an additional capability to inquire the upstream node of its upstream node, the procedure can iteratively be applied until the fault is localized or the root node is reached. In all cases the TTL for the request need only be at most 2. Thus the processing load of each request is small as only a limited number of nodes will receive the request.

This document defines protocol extensions to MPLS ping [RFC4379] to allow a third party to remotely cause an MPLS echo request message to be sent down a Label Switched Path (LSP) or part of an LSP. The procedure described in the paragraphs above does require that the initiator know the previous-hop node to the one which was pinged on the prior iteration. This information is readily available in [RFC4875]. This document also provides a means for obtaining this information for [RFC6388].

While the motivation for this document came from multicast scaling concerns, it's applicability may be wider. However other uses of this facility are beyond the scope of this document. In particular, the procedures defined in this document only allow testing of a FEC stack consisting of a single FEC. It also does not allow the initiator to specify the label assigned to that FEC, nor does it allow the initiator to cause any additional labels to be added to the label stack of the actual MPLS echo request message.

1.1. Requirements Language

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

The term "Must Be Zero" (MBZ) is used in object descriptions for reserved fields. These fields MUST be set to zero when sent and ignored on receipt.

Based on context the terms leaf and egress are used interchangeably. Egress is used where consistency with [RFC4379] was deemed appropriate. Receiver is used in the context of receiving protocol messages.

[Note (to be removed after assignments occur): <tba> = to be assigned by IANA]

2. Proxy Ping Overview

This document defines a protocol interaction between a first node and a node which is part of an LSP to allow the first node to request that second node initiate an LSP ping for the LSP on behalf of the first node. Two new LSP Ping messages are defined for remote pinging: the MPLS proxy ping request and the MPLS proxy ping reply.

A remote ping operation on a P2MP LSP generally involves at least three LSRs; in some scenarios none of these are the ingress (root) or an egress (leaf) of the LSP.

We refer to these nodes with the following terms:

Initiator - the node which initiates the ping operation by sending an MPLS proxy ping request message

Proxy LSR - the node which is the destination of the MPLS proxy request message and potential initiator of the MPLS echo request

Receiver(s) - the nodes which receive the MPLS echo request message

Responder - A receiver that responds to a MPLS Proxy Ping Request or an MPLS Echo Request

We note that in some scenarios, the initiator could also be the responder, in which case the response would be internal to the node.

The initiator formats an MPLS proxy ping request message and sends it to the proxy LSR, a node it believes to be on the path of the LSP. This message instructs the proxy LSR to either Reply with Proxy information or to send a MPLS echo request inband of the LSP. The initiator requests Proxy information so that it can learn additional information it needs to use to form a subsequent MPLS Proxy Ping request. For example during LSP traceroute an initiator needs the downstream map information to form an Echo request. An initiator may also want to learn a Proxy LSR's FEC neighbor information so that it can form proxy request to various nodes along the LSP.

The proxy LSR either replies with the requested Proxy information or it validates that it has a label mapping for the specified FEC and that it is authorized to send the specified MPLS echo request on behalf of the initiator.

If the proxy LSR has a label mapping for the FEC and all authorization checks have passed, the proxy LSR formats an MPLS echo request. If the source address of the MPLS echo request is not to be set to the Proxy Request source address, the initiator must include a Reply-to Address object containing the source address to use in the MPLS echo request. It then sends it inband of the LSP.

The receivers process the MPLS echo request as normal, sending their MPLS echo replies back to the initiator.

If the proxy LSR failed to send a MPLS echo request as normal because it encountered an issue while attempting to send, a MPLS proxy ping reply message is sent back with a return code indicating that the MPLS echo request could not be sent.

3. Proxy MPLS Echo Request / Reply Procedures

3.1. Procedures for the initiator

The initiator creates an MPLS proxy ping request message.

The message MUST contain a Target FEC Stack that describes the FEC being tested. The topmost FEC in the target FEC stack is used at the Proxy LSR to lookup the MPLS label stack that will be used to encapsulate the MPLS echo request packet.

The MPLS Proxy Ping message MUST contain a Proxy Echo Parameters object. In that object, the address type is set to either IPv4 or IPv6. The Destination IP Address is set to the value to be used in the MPLS echo request packet. If the Address Type is IPv4, an address is from the range 127/8. If the Address Type is IPv6, an address is from the range ::FFFF:7F00:0/104.

The Reply mode and Global Flags of the Proxy Echo Parameters object are set to the values to be used in the MPLS echo request message header. The Source UDP Port is set to the value to be used in the MPLS echo request packet. The TTL is set to the value to be used in the outgoing MPLS label stack. See Section 5.1 for further details.

If the FEC's Upstream/Downstream Neighbor address information is required, the initiator sets the "Request for FEC neighbor information" Proxy Flags in the Proxy Echo Parameters object.

If a Downstream Detailed or Downstream Mapping TLV is required in a MPLS Proxy Ping Reply, the initiator sets the "Request for Downstream Detailed Mapping" or "Request for Downstream Mapping" Proxy Flags in the Proxy Echo Parameters object. Only one of the two flags can be set.

The Proxy Request reply mode is set with one of the reply modes defined in [RFC4379] as appropriate.

A list of Next Hop IP Addresses MAY be included to limit the next hops towards which the MPLS echo request message will be sent. These are encoded as Next Hop sub-objects and included in the Proxy Echo Parameters object.

Proxy Echo Parameter object MPLS payload size field may be set to request that the MPLS echo request (including any IP and UDP header) be zero padded to the specified size. When the payload size is non zero, if sending the MPLS Echo Request involves using an IP header, the DF bit MUST be set to 1.

Any of following objects MAY be included; these objects will be copied into the MPLS echo request messages:

Pad

Vendor Enterprise Number

Reply TOS Byte

P2MP Responder Identifier [RFC6425]

Echo Jitter TLV [RFC6425]

Vendor Private TLVs

Downstream Detailed Mapping or Downstream Mapping objects MAY be included. These objects will be matched to the next hop address for inclusion in those particular MPLS echo request messages.

The message is then encapsulated in a UDP packet. The source UDP port is chosen by the initiator; the destination UDP port is set to 3503. The IP header is set as follows: the source IP address is a routable address of the initiator; the destination IP address is a routable address to the Proxy LSR. The packet is then sent with the IP TTL is set to 255.

3.2. Procedures for the proxy LSR

A proxy LSR that receives an MPLS proxy ping request message, parses the packet to ensure that it is a well-formed packet. It checks that the TLVs that are not marked "Ignore" are understood. If not, it sets the Return Code set to "Malformed echo request received" or "TLV not understood" (as appropriate), and the Subcode set to zero. If the Reply Mode of the message header is not 1(Do not reply), an MPLS proxy ping reply message SHOULD be sent as described below. In the latter case, the misunderstood TLVs (only) are included in an Errored TLVs object.

The Proxy LSR checks that the MPLS proxy ping request message did not arrive via one of its exception processing paths. Packets arriving via IP TTL expiry, IP destination address set to a Martian address or label ttl expiry MUST be treated as "Unauthorized" packets. An MPLS proxy ping reply message MAY be sent with a Return Code of <tba>, "Proxy Ping not authorized".

The header fields Sender's Handle and Sequence Number are not examined, but are saved to be included in the MPLS proxy ping reply or MPLS echo request messages.

The proxy LSR validates that it has a label mapping for the specified FEC, it then determines if it is an ingress, egress, transit or bud node and sets the Return Code as appropriate. A new return code (Replying router has FEC mapping for topmost FEC) has been defined for the case where the Proxy LSR is an ingress (for example head of the TE tunnel or a transit router) because the existing RFC4379 return codes don't match the situation. For example, when a Proxy LSR is a transit router, it's not appropriate for the return code to describe how the packet would transit because the Proxy Request doesn't contain information about what input interface the an MPLS echo request would be switched from at the Proxy LSR.

The proxy LSR then determines if it is authorized to send the specified MPLS echo request on behalf of the initiator. A Proxy LSR MUST be capable of filtering addresses to validate initiators. Other filters on FECs or MPLS echo request contents MAY be applied. If a filter has been invoked (i.e. configured) and an address does not pass the filter, then an MPLS echo request message MUST NOT be sent, and the event SHOULD be logged. An MPLS proxy ping reply message MAY be sent with a Return Code of <tba>, "Proxy Ping not authorized".

The destination address specified in the Proxy Echo Parameters object is checked to ensure that it conforms to the address allowed IPv4 or IPv6 address range. If not, it sets the Return Code set to "Malformed echo request received" and the Subcode set to zero. If the Reply Mode of the message header is not 1, an MPLS proxy ping reply message SHOULD be sent as described below.

If the "Request for FEC Neighbor Address info" flag is set, a Upstream Neighbor Address Object and/or Downstream Neighbor Address Object(s) is/are formatted for inclusion in the MPLS proxy ping reply. If the Upstream or Downstream address is unknown they are not included in the Proxy Reply.

If there are Next Hop sub-objects in the Proxy Echo Parameters object, each address is examined to determine if it is a valid next hop for this FEC. If any are not, Proxy Echo Parameters object should be updated removing unrecognized Next Hop sub-objects. The updated Proxy Echo Parameters object MUST be included in the MPLS proxy ping reply.

If the "Request for Downstream Detailed Mapping" or "Request for Downstream Mapping" flag is set, the LSR formats (for inclusions in the MPLS proxy ping reply) a Downstream Detailed/Downstream Mapping object for each interface over which the MPLS echo request will be sent.

If the Proxy LSR is the egress for the FEC, the behavior of the proxy

LSR vary depending on whether the node is an Egress of a P2P LSP, a P2MP LSP or MP2MP LSP. Additional details can be found in the section describing "Handling when Proxy LSR it is egress for FEC".

If the Reply Mode of the Proxy Request message header is "1 - do not reply", no MPLS proxy ping reply is sent. Otherwise an MPLS proxy ping reply message or MPLS echo request should be sent as described below.

3.2.1. Proxy LSR Handling when it is Egress for FEC

This sections describes the different behaviors for the Proxy LSR when it's the Egress for the FEC. In the P2MP budnode and MP2MP budnode and egress cases, different behavior is required.

When the Proxy LSR is the egress of a P2P FEC, a Proxy reply should be sent to the initiator with the return code set to 3 (Reply router is Egress for FEC) with return subcode set to 0.

When the Proxy LSR is the egress of a P2MP FEC, it can be either a budnode or just an Egress. If the Proxy LSR is a Budnode, a Proxy reply should be sent to the initiator with the return code set to 3 (Reply router is Egress for FEC) with return subcode set to 0 and DS/DDMAPs only if the Proxy initiator requested information to be returned in a Proxy reply. If the Proxy LSR is a Budnode but not requested to return a Proxy reply, the Proxy LSR should send packets to the downstream neighbors (no Echo reply is sent to the Proxy Initiator to indicate that the Proxy LSR is an egress). If the Proxy LSR is just an egress, a Proxy reply should be sent to the initiator with the return code set to 3 (Reply router is Egress for FEC) with return subcode set to 0.

When the Proxy LSR is the egress of a MP2MP FEC, it can be either a budnode or just an Egress. LSP pings sent from a leaf of a MP2MP has different behavior in this case. MPLS echo request are sent to all upstream/downstream neighbors. The Proxy LSRs need to be consistent with this variation in behavior. If the Proxy LSR is a Budnode or just an egress, a Proxy reply should be sent to the initiator with the return code set to 3 (Reply router is Egress for FEC) with return subcode set to 0 and DS/DDMAPs included only if the Proxy initiator requested information to be returned in a Proxy reply. If the Proxy LSR is not requested to return information in a proxy reply, the Proxy LSR should send packets to all upstream/downstream neighbors as would be done when sourcing an LSP ping from a M2MP leaf (no echo reply is sent to the Proxy initiator indicating that the Proxy LSR is an egress).

3.2.2. Downstream Detailed/Downstream Maps in Proxy Reply

When the Proxy LSR is a transit or bud node, downstream maps corresponding to how the packet is transited can not be supplied unless an ingress interface for the MPLS echo request is specified, since this information is not available and since all valid output paths are of interest, the Proxy LSR should include DS/DDMAP(s) to describe the entire set of paths that the packet can be replicated, like in the case where an LSP ping is initiated at the Proxy LSR. For mLDP there is a DMAP/DDMAP per upstream/downstream neighbor for MP2MP LSPs, or per downstream neighbor in the P2MP LSP case.

When the Proxy LSR is a bud node or egress in a MP2MP LSP or a budnode in a P2MP LSP, an LSP ping initiated from the Proxy LSR would source packets only to the neighbors but not itself despite the fact that the Proxy LSR is itself an egress for the FEC. In order to match the behavior as seen from LSP Ping initiated at the Proxy LSR, the Proxy Reply should contain DMAP/DDMAPs for only the paths to the upstream/downstream neighbors, but no DMAP/DDMAP describing its own egresses paths. The proxy LSR identifies that it's an egress for the FEC using a different Proxy Reply return code. The Proxy reply return code is either set to "Reply router has a mapping for the topmost FEC" or "Reply router is Egress for the FEC".

3.2.3. Sending an MPLS proxy ping reply

The Reply mode, Sender's Handle and Sequence Number fields are copied from the proxy ping request message. The objects specified above are included. The message is encapsulated in a UDP packet. The source IP address is a routable address of the proxy LSR; the source port is the well-known UDP port for LSP ping. The destination IP address and UDP port are copied from the source IP address and UDP port of the echo request. The IP TTL is set to 255.

3.2.4. Sending the MPLS echo requests

A base MPLS echo request is formed as described in the next section. The section below that describes how the base MPLS echo request is sent on each interface.

3.2.4.1. Forming the base MPLS echo request

A Next_Hop_List is created as follows. If Next Hop sub-objects were included in the received Proxy Parameters object, the Next_Hop_List created from the address in those sub-objects as adjusted above. Otherwise, the list is set to all the next hops to which the FEC would be forwarded.

The proxy LSR then formats an MPLS echo request message. The Global Flags and Reply Mode are copied from the Proxy Echo Parameters object. The Return Code and Return Subcode are set to zero.

The Sender's Handle and Sequence Number are copied from the remote echo request message.

The TimeStamp Sent is set to the time-of-day (in seconds and microseconds) that the echo request is sent. The TimeStamp Received is set to zero.

If the reply-to address object is present, it is used to set the echo request source address, otherwise the echo request source address is set to the proxy request source address.

The following objects are copied from the MPLS proxy ping request message. Note that of these, only the Target FEC Stack is REQUIRED to appear in the MPLS proxy ping request message.

Target FEC Stack

Pad

Vendor Enterprise Number

Reply TOS Byte

P2MP Responder Identifier [RFC6425]

Echo Jitter TLV [RFC6425]

Vendor Private TLVs

The message is then encapsulated in a UDP packet. The source UDP port is copied from the Proxy Echo Parameters object. The destination port copied from the proxy ping request message.

The source IP address is set to a routable address specified in the reply-to-address object or the source address of the received proxy request. Per usual the TTL of the IP packet is set to 1.

If the Explicit DSCP flag is set, the Requested DSCP byte is examined. If the setting is permitted then the DSCP byte of the IP header of the MPLS Echo Request message is set to that value. If the Proxy LSR does not permit explicit control for the DSCP byte, the MPLS Proxy Echo Parameters with the Explicit DSCP flag cleared MUST be included in any MPLS proxy ping reply message to indicate why an Echo Request was not sent. The return code MUST be set to <tba>.

"Proxy ping parameters need to be modified". If the Explicit DSCP flag is not set, the Proxy LSR should set the Echo Request DSCP settings to the value normally used to source LSP ping packets..

3.2.4.2. Per interface sending procedures

The proxy LSR now iterates through the Next_Hop_List modifying the base MPLS echo request to form the MPLS echo request packet which is then sent on that particular interface.

For each next hop address, the outgoing label stack is determined. The TTL for the label corresponding to the FEC specified in the FEC stack is set such that the TTL on the wire will be the TTL specified in the Proxy Echo Parameters. If any additional labels are pushed onto the stack, their TTLs are set to 255.

If the MPLS proxy ping request message contained Downstream Mapping/Downstream Detailed Mapping objects, they are examined. If the Downstream IP Address matches the next hop address that Downstream Mapping object is included in the MPLS echo request.

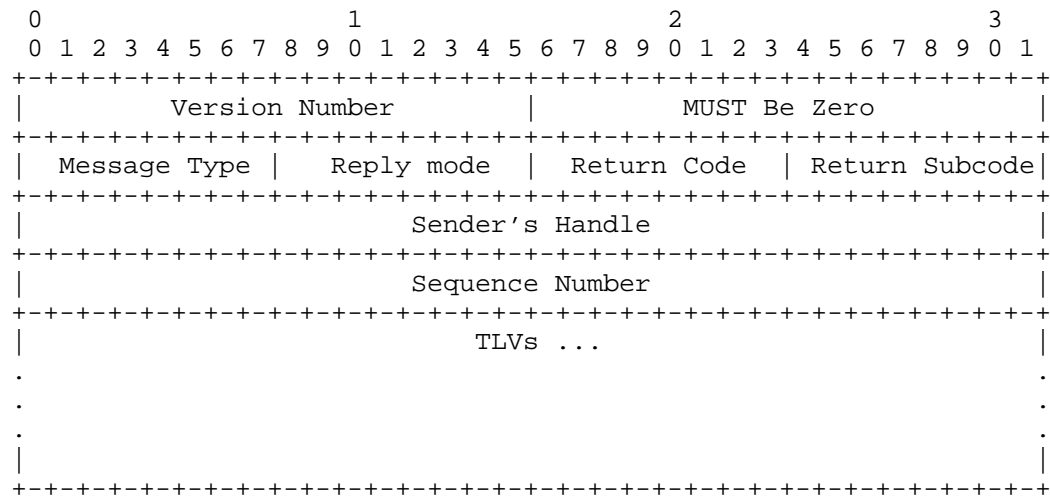
The packet is then transmitted on this interface.

4. Proxy Ping Request / Reply Messages

This document defines two new LSP Ping messages, the MPLS proxy ping request and the MPLS proxy ping reply.

4.1. Proxy Ping Request / Reply Message formats

Except where noted, the definitions of all fields in the messages are identical to those found in [RFC4379]. The messages have the following format:



Version Number

The Version Number is currently 1. (Note: the Version Number is to be incremented whenever a change is made that affects the ability of an implementation to correctly parse or process an MPLS echo request/reply. These changes include any syntactic or semantic changes made to any of the fixed fields, or to any TLV or sub-TLV assignment or format that is defined at a certain version number. The Version Number may not need to be changed if an optional TLV or sub-TLV is added.)

Message Type

| Type | Message |
|------|--|
| ---- | ----- |
| 3 | MPLS proxy ping request
(Pending IANA assignment) |
| 4 | MPLS proxy ping reply
(Pending IANA assignment) |

4.2. Proxy Ping Request Message contents

The MPLS proxy ping request message MAY contain the following objects:

| Type | Object |
|------|---|
| ---- | ----- |
| 1 | Target FEC Stack |
| 2 | Downstream Mapping |
| 3 | Pad |
| 5 | Vendor Enterprise Number |
| 10 | Reply TOS Byte |
| 11 | P2MP Responder Identifier [RFC6425] |
| 12 | Echo Jitter TLV [RFC6425] |
| 20 | Downstream Detailed Mapping |
| 30 | Proxy Echo Parameters (Pending IANA assignment) |
| * | Vendor Private TLVs |

* TLVs types in the Vendor Private TLV Space MUST be ignored if not understood

4.3. Proxy Ping Reply Message Contents

The MPLS proxy ping reply message MAY contain the following objects:

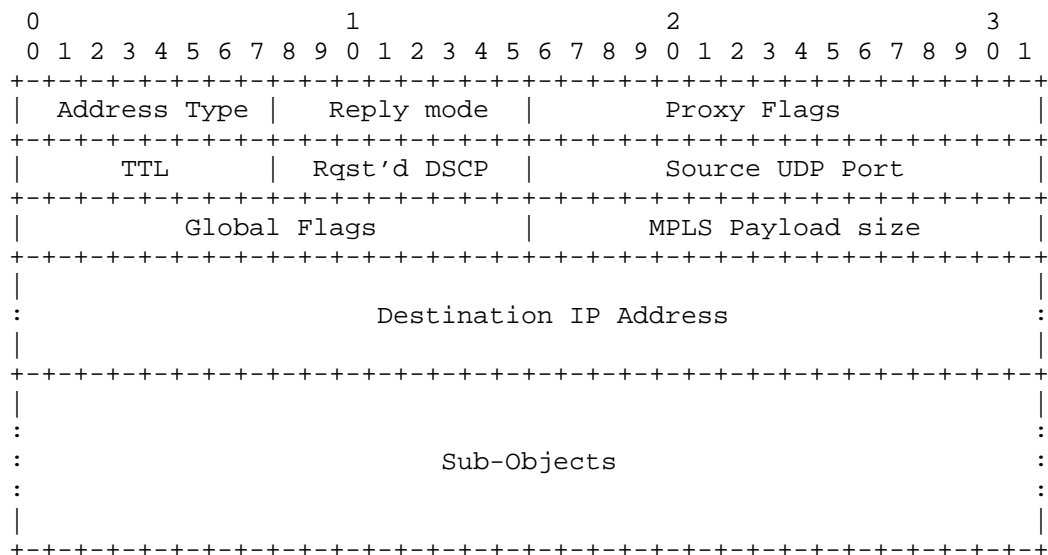
| Type | Object |
|------|--|
| ---- | ----- |
| 1 | Target FEC Stack |
| 2 | Downstream Mapping |
| 5 | Vendor Enterprise Number |
| 9 | Errored TLVs |
| 20 | Downstream Detailed Mapping |
| 30 | Proxy Echo Parameters
(Pending IANA assignment) |
| 31 | Upstream Neighbor Address |
| 32 | Downstream Neighbor Address (0 or more) |
| * | Vendor Private TLVs |

* TLVs types in the Vendor Private TLV Space MUST be ignored if not understood

5. Object formats

5.1. Proxy Echo Parameters Object

The Proxy Echo Parameters object is a TLV that MUST be included in an MPLS Proxy Echo Request message. The length of the TLV is $12 + K + S$, where K is the length of the Destination IP Address field and S is the total length of the sub-objects. The Proxy Echo Parameters object can be used to either to 1) control attributes used in Composing and Sending an MPLS echo request or 2) query the Proxy LSR for information about the topmost FEC in the target FEC stack but not both. In the case where the Proxy LSR is being queried (ie information needs to be returned in a Proxy Reply), no MPLS echo request will be sent from the Proxy LSR. The MPLS Proxy Echo request echo header's Reply Mode should be set to "Reply with Proxy Info".



Address Type

The type and length of the address found in the in the Destination IP Address and Next Hop IP Addresses fields. The type codes appear in the table below:

| Address Family | Type | Length |
|----------------|------|--------|
| IPv4 | 1 | 4 |
| IPv6 | 3 | 16 |

Reply mode

The reply mode to be sent in the MPLS Echo Request message; the

values are as specified in [RFC4379].

Proxy Flags

The Proxy Request Initiator sets zero, one or more of these flags to request actions at the Proxy LSR.

Request for FEC Neighbor Address info 0x01

When set this requests that the proxy LSR supply the Upstream and Downstream neighbor address information in the MPLS proxy ping reply message. This flag is only applicable for the topmost FEC in the FEC stack if the FEC types corresponds with a P2MP or MP2MP LSPs. The Proxy LSR MUST respond as applicable with a Upstream Neighbor Address Object and Downstream Neighbor Address Object(s) in the MPLS Proxy ping reply message. Upstream Neighbor Address Object needs be included only if there is an upstream neighbor. Similarly, one Downstream Neighbor Address Object needs to be included for each Downstream Neighbor for which the LSR learned bindings from.

Setting this flag will cause the proxy LSR to cancel sending an Echo request. Information learned with such proxy reply may be used by the proxy initiator to generate subsequent proxy requests.

Request for Downstream Mapping 0x02

When set this requests that the proxy LSR supply a Downstream Mapping object see [RFC4379] in the MPLS proxy ping reply message. It's not valid to have Request for Downstream Detailed Mapping flag set when this flag is set.

Setting this flag will cause the proxy LSR to cancel sending an Echo request. Information learned with such proxy reply may be used by the proxy initiator to generate subsequent proxy requests.

Request for Downstream Detailed Mapping 0x04

When set this requests that the proxy LSR supply a Downstream Detailed Mapping object see [RFC6424] in the MPLS proxy ping reply message. It's not valid to have Request for Downstream Mapping flag set when this flag is set.

Setting this flag will cause the proxy LSR to cancel sending

an Echo request. Information learned with such proxy reply may be used by the proxy initiator to generate subsequent proxy requests.

Explicit DSCP Request 0x08

When set this requests that the proxy LSR use the supplied "Rqst'd DSCP" byte in the echo request message

TTL

The TTL to be used in the label stack entry corresponding to the topmost FEC in the in the MPLS Echo Request packet. Valid values are in the range [1,255]. A setting of 0 should be ignored by the Proxy LSR.

Requested DSCP

This field is valid only if the Explicit DSCP flag is set. If not set, the field MUST be zero on transmission and ignored on receipt. When the flag is set this field contains the DSCP value to be used in the MPLS echo request packet IP header.

Source UDP Port

The source UDP port to be sent in the MPLS Echo Request packet

Global Flags

The Global Flags to be sent in the MPLS Echo Request message

MPLS Payload Size

Used to request that the MPLS payload (IP header + UDP header + MPLS echo request) be padded using a zero filled Pad TLV so that the IP header, UDP header nad MPLS echo request total the specified size. Field set to zero means no size request is being made. If the requested size is less than the minimum size required to form the MPLS echo request, the request will be treated as a best effort request with the Proxy LSR building the smallest possible packet (ie not using a Pad TLV). The IP header DF bit should be set when this field is non zero.

Destination IP Address

If the Address Type is IPv4, an address from the range 127/8;

If the Address Type is IPv6, an address from the range
::FFFF:7F00:0/104

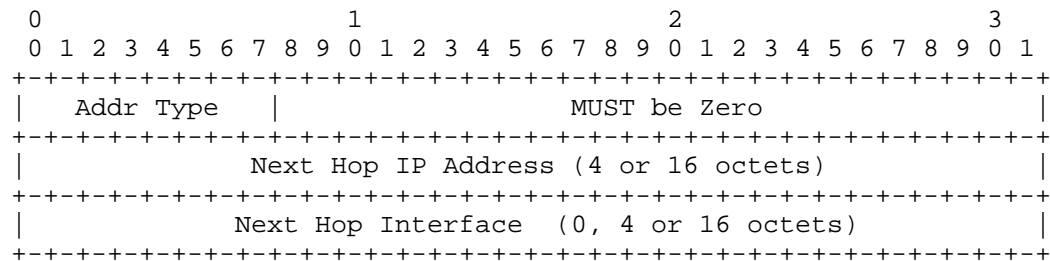
Sub-Objects

A TLV encoded list of sub-objects. Currently one is defined.

| Sub-Type | Length | Value Field |
|----------|--------|-------------|
| ----- | ----- | ----- |
| 1 | 8+ | Next Hop |

5.1.1.1. Next Hop sub-Object

This sub-object is used to describe a particular next hop towards which the Echo Request packet should be sent. If the topmost FEC in the FEC-stack is a multipoint LSP, this sub-object may appear multiple times.



Address Type

| Type | Type of Next Hop | Addr Length | IF Length |
|------|-------------------|-------------|-----------|
| 1 | IPv4 Numbered | 4 | 4 |
| 2 | IPv4 Unnumbered | 4 | 4 |
| 3 | IPv6 Numbered | 16 | 16 |
| 4 | IPv6 Unnumbered | 16 | 4 |
| 5 | IPv4 Protocol Adj | 4 | 0 |
| 6 | IPv6 Protocol Adj | 16 | 0 |

Note: Types 1-4 correspond to the types in the DS Mapping object. They are expected to be populated with information obtained through a previously returned DS Mapping object. Types 5 and 6 are intended to be populated from the local address information obtained from a previously returned Previous Hop Address Object.

Next Hop IP Address

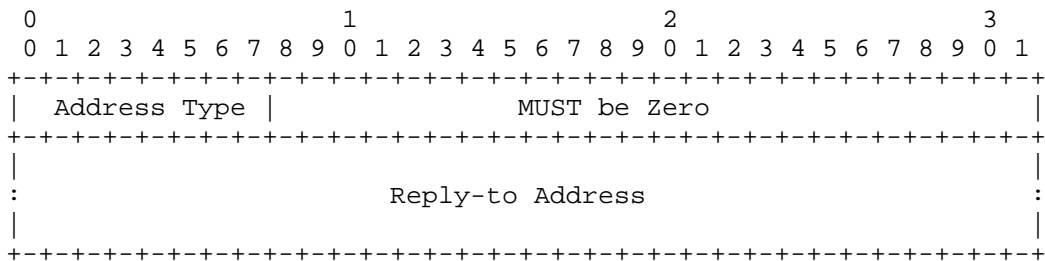
A next hop address that the echo request message is to be sent towards

Next Hop Interface

Identifier of the interface through which the echo request message is to be sent

5.2. Reply-to Address Object

Used to specify the MPLS echo request IP source address. This address must be IP reachable via the Proxy LSR otherwise it will be rejected.

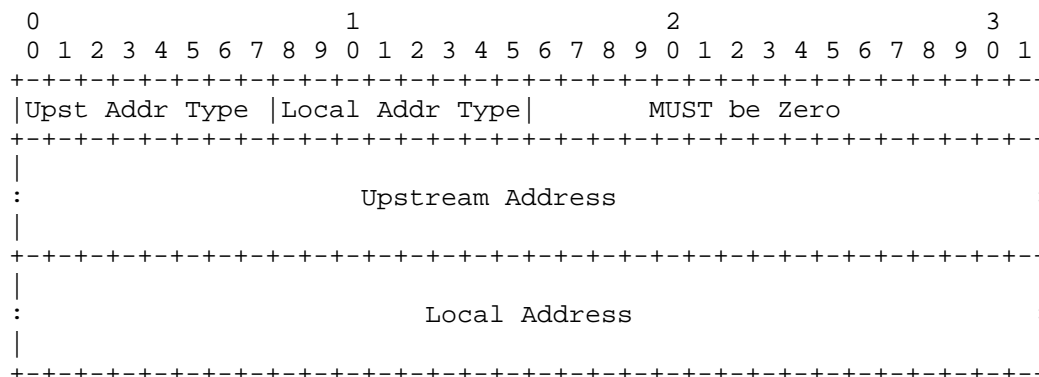


Address Type

A type code as specified in the table below:

| Type | Type of Address |
|------|-----------------|
| 1 | IPv4 |
| 3 | IPv6 |

5.3. Upstream Neighbor Address Object



Upst Addr Type; Local Addr Type

These two fields determine the type and length of the respective addresses. The codes are specified in the table below:

| Type | Type of Address | Length |
|------|---------------------|--------|
| 0 | No Address Supplied | 0 |
| 1 | IPv4 | 4 |
| 3 | IPv6 | 16 |

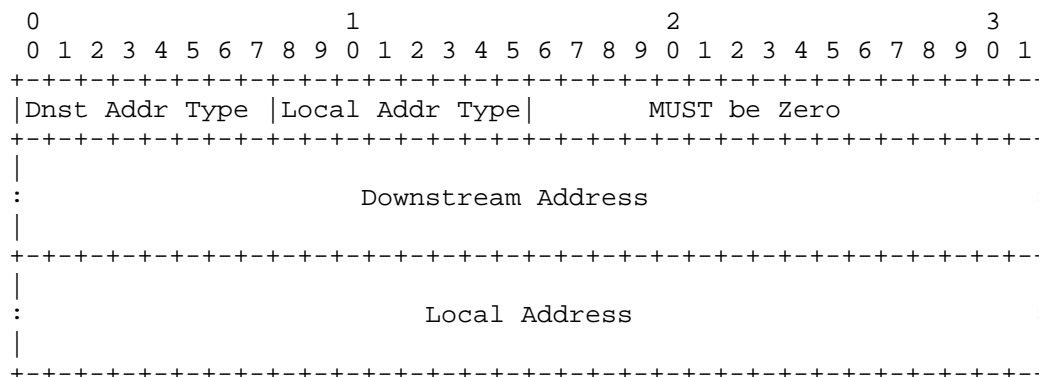
Upstream Address

The address of the immediate upstream neighbor for the topmost FEC in the FEC stack. If protocol adjacency exists by which the label for this FEC was exchanged, this address MUST be the address used in that protocol exchange.

Local Address

The local address used in the protocol adjacency exists by which the label for this FEC was exchanged.

5.4. Downstream Neighbor Address Object



Dnst Addr Type; Local Addr Type

These two fields determine the type and length of the respective addresses. The codes are specified in the table below:

| Type | Type of Address | Length |
|------|---------------------|--------|
| 0 | No Address Supplied | 0 |
| 1 | IPv4 | 4 |
| 3 | IPv6 | 16 |

Downstream Address

The address of a immediate downstream neighbor for the topmost FEC in the FEC stack. If protocol adjacency exists by which the label for this FEC was exchanged, this address MUST be the address used in that protocol exchange.

Local Address

The local address used in the protocol adjacency exists by which the label for this FEC was exchanged.

6. Security Considerations

The mechanisms described in this document are intended to be used within a Service Provider network and to be initiated only under the authority of that administration.

If such a network also carries internet traffic, or permits IP access from other administrations, MPLS proxy ping message SHOULD be

discarded at those points. This can be accomplished by filtering on source address or by filtering all MPLS ping messages on UDP port.

Any node which acts as a proxy node SHOULD validate requests against a set of valid source addresses. An implementation MUST provide such filtering capabilities.

MPLS proxy ping request messages are IP addressed directly to the Proxy node. If a node which receives an MPLS proxy ping message via IP or Label TTL expiration, it MUST NOT be acted upon.

MPLS proxy ping request messages are IP addressed directly to the Proxy node. If a MPLS Proxy ping request IP destination address is a Martian Address, it MUST NOT be acted upon.

if a MPLS Proxy ping request IP source address is not IP reachable by the Proxy LSR, the Proxy request MUST NOT be acted upon.

MPLS proxy ping requests are limited to making their request via the specification of a FEC. This ensures that only valid MPLS echo request messages can be created. No label spoofing attacks are possible.

7. Acknowledgements

The authors would like to thank Nobo Akiya for his detailed review and insightful comments.

8. IANA Considerations

This document makes the following assignments (pending IANA action)

LSP Ping Message Types

| Type | Value Field |
|---------|-------------------------|
| ---- | ----- |
| 03(tba) | MPLS proxy ping request |
| 04(tba) | MPLS proxy ping reply |

Objects and Sub-Objects

| Type | Sub-Type | Value Field |
|---------|----------|-----------------------------|
| ---- | ----- | ----- |
| 22(tba) | 1 | Proxy Echo Parameters |
| | | Next Hop |
| 23(tba) | | Reply-to Address |
| 24(tba) | | Upstream Neighbor Address |
| 25(tba) | | Downstream Neighbor Address |

Return Code [pending IANA assignment]

| Value | Meaning |
|---------|--|
| ----- | ----- |
| 16(tba) | Proxy ping not authorized. |
| 17(tba) | Proxy ping parameters need to be modified. |
| 18(tba) | MPLS Echo Request Could not be sent. |
| 18(tba) | Replying router has FEC mapping for topmost FEC. |

9. References

9.1. Normative References

- [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.

9.2. Informative References

- [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.
- [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.

Authors' Addresses

George Swallow
Cisco Systems
1414 Massachusetts Ave
Boxborough, MA 01719
USA

Email: swallow@cisco.com

Vanson Lim
Cisco Systems
1414 Massachusetts Avenue
Boxborough, MA 01719
USA

Email: vlim@cisco.com

Sam Aldrin
Huawei Technologies
2330 Central Express Way
Santa Clara, CA 95951
USA

Email: aldrin.ietf@gmail.com

Network Working Group
Internet-Draft
Intended status: Experimental
Expires: August 17, 2013

E. Osborne
Cisco
February 13, 2013

Extended Administrative Groups in MPLS-TE
draft-osborne-mpls-extended-admin-groups-00

Abstract

This document provides additional administrative groups (sometimes referred to as "link colors") to the IGP extensions for MPLS-TE.

Requirements Language

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

Status of this Memo

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

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

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

This Internet-Draft will expire on August 17, 2013.

Copyright Notice

Copyright (c) 2013 IETF Trust and the persons identified as the document authors. All rights reserved.

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

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

Table of Contents

| | |
|---|---|
| 1. Introduction | 3 |
| 2. Extended Administrative Groups sub-TLV | 3 |
| 2.1. Packet Format | 3 |
| 2.2. Admin group numbering | 4 |
| 2.3. Backward compatability | 4 |
| 2.3.1. AG and EAG coexistence | 4 |
| 2.3.2. Desire for unadvertised EAG bits | 4 |
| 3. Security Considerations | 4 |
| 4. IANA Considerations | 5 |
| 5. Acknowledgements | 5 |
| 6. Normative References | 5 |
| Author's Address | 5 |

1. Introduction

MPLS-TE advertises 32 administrative groups (commonly referred to as "colors" or "link colors") using the Administrative Group sub-TLV of the Link TLV. This is defined for OSPF [RFC3630] and ISIS [RFC5305].

This document adds a sub-TLV to the IGP TE extensions, "Extended Administrative Group". It

2. Extended Administrative Groups sub-TLV

The Extended Administrative Groups sub-TLV is used in addition to the Administrative Groups when a device wishes to advertise more than 32 colors for a link. The EAG sub-TLV is optional.

This document uses the term 'colors' as a shorthand to refer to particular bits with an AG or EAG. The examples in this document use 'red' to represent the least significant bit in the AG (red == 0x1), 'blue' to represent the second bit (blue == 0x2). To say that a link has a given color or that the specified color is set on the link is to say that the corresponding bit or bits in the link's AG are set to 1.

2.1. Packet Format

The format of the Extended Administrative Groups sub-TLV is:

```

      0               1               2               3
      0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-----+-----+-----+-----+-----+-----+-----+-----+
| Type: Extended Admin Group | Length: Variable |
+-----+-----+-----+-----+-----+-----+-----+-----+
| Value: Extended Admin Group Value |
+-----+-----+-----+-----+-----+-----+-----+-----+
|                                     .....
+-----+-----+-----+-----+-----+-----+-----+-----+
| Value: Extend Admin Group Value |
+-----+-----+-----+-----+-----+-----+-----+-----+

```

The Length is the size of the Extended Admin Group (EAG) value in octets. The EAG may be of any length, but must be a multiple of 4 octets.

2.2. Admin group numbering

By convention, the existing Administrative Group TLVs are numbered 0 (LSB) to 31 (MSB). The EAG values pick up where this numbering scheme leaves off. The LSB in the EAG is 32. If the EAG is 4 bytes in length, the MSB is 63. If the EAG is 8 bytes in length, the MSB is 95.

2.3. Backward compatability

There are two things to consider for backward compatibility with existing AG implementations - how do AG and EAG coexist, and what happens if a node has matching criteria for unadvertised EAG bits?

2.3.1. AG and EAG coexistence

If a node advertises the EAG sub-TLV it MUST also advertise the existing Administrative Group (AG) sub-TLV defined in RFCs 3630 and 5305. This ensures that the first bit of the EAG sub-TLV is always bit 32, and ensures maximum interoperability with legacy implementations.

2.3.2. Desire for unadvertised EAG bits

The existing AG bits are optional; thus a node may be configured with a preference to include red or exclude blue, and be faced with a link that is not advertising a value for either blue or red. What does an implementation do in this case? It shouldn't assume that red is set, but it is also arguably incorrect to assume that red is NOT set, as a bit must first exist before it can be set to 0.

Practically speaking this has not been an issue for deployments, as many implementations always advertise the AG bits, often with a default value of 0x00000000. However, this issue may be of more concern once EAGs are added to the network. EAGs may exist on some nodes but not others, and the EAG length may be longer for some links than for others.

Each implementation is free to choose its own method for handling this question. However, to encourage maximum interoperability an implementation SHOULD treat specified but unadvertised EAG bits as if they are set to 0. A node MAY provide other (configurable) strategies for handling this case.

3. Security Considerations

This extension adds no new security considerations.

4. IANA Considerations

This document requests a sub-TLV allocation in both OSPF and ISIS.

5. Acknowledgements

Thanks to Santiago Alvarez and Rohit Gupta for their review and comments.

6. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC3630] Katz, D., Kompella, K., and D. Yeung, "Traffic Engineering (TE) Extensions to OSPF Version 2", RFC 3630, September 2003.
- [RFC5305] Li, T. and H. Smit, "IS-IS Extensions for Traffic Engineering", RFC 5305, October 2008.

Author's Address

Eric Osborne
Cisco

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: August 15, 2013

E. Osborne
February 11, 2013

Updates to PSC
draft-osborne-mpls-psc-updates-00

Abstract

This document contains four updates to RFC6378, "MPLS Transport Profile (MPLS-TP) Linear Protection". Two of them correct existing behavior. The third clears up a behavior which was not explained in the RFC, and the fourth adds rules around handling capabilities mismatches.

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 August 15, 2013.

Copyright Notice

Copyright (c) 2013 IETF Trust and the persons identified as the document authors. All rights reserved.

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

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

Table of Contents

| | |
|---|---|
| 1. Introduction | 3 |
| 2. Incorrect local status after failure | 3 |
| 3. Reversion deadlock due to a race condition | 3 |
| 4. Clarifying PSC's behavior in the face of multiple inputs | 4 |
| 5. Security Considerations | 6 |
| 6. IANA Considerations | 6 |
| 7. Acknowledgements | 7 |
| 8. References | 7 |
| 8.1. Normative References | 7 |
| 8.2. Informative References | 7 |
| Author's Address | 7 |

1. Introduction

This document contains four updates to PSC [RFC6378] . Three of them fix issues identified in the ITU's liaison statement "Recommendation ITU-T G.8131/Y.1382 revision - Linear protection switching for MPLS-TP networks" [LIAISON]. The fourth clears up a behavior which was not well explained in RFC6378. These updates are not changes to the protocol's packet format or to PSC's design, but are corrections and clarifications to specific aspects of the protocol's procedures..

2. Incorrect local status after failure

Issue #2 in the liaison identifies a case where a strict reading of RFC6378 leaves a node reporting an inaccurate status

. A node can end up sending incorrect status - NR(0,1) - despite the failure of the protection LSP (P-LSP). This is clearly not correct, as a node should not be sending NR if it has a local failure. To address this issue, the fourth bullet in section 4.3.3.3 is replaced with the following three bullets:

- o If the current state is due to a local or remote Manual Switch, a local Signal Fail indication on the protection path SHALL cause the LER to enter local Unavailable state and begin transmission of an SF(0,0) message.
- o If the LER is in local Protecting Administrative state due to a local Forced Switch, a local Signal Fail indication on the protection path SHALL be ignored.
- o If the LER is in remote Protecting Administrative state due to a remote Forced Switch, a local Signal Fail indication on the protection path SHALL cause the LER to remain in remote Protecting administrative state and transmit an SF(0,1) message.

3. Reversion deadlock due to a race condition

Issue #8 in the liaison identifies a deadlock case where each node can end up sending NR(0,1) when it should instead be in the process of recovering from the failure (i.e. entering into WTR or DNR, as appropriate for the protection domain). The root of the issue is that a pair of nodes can simultaneously enter WTR state, receive an out of date SF-W indication and transition into a remotely triggered WTR, and remain in remotely triggered WTR waiting for the other end to trigger a change in status.

In the case identified in issue #8, each node can end up sending NR(0,1), which is an indication that the transmitting node has no local failure, but is instead reacting to the remote SF-W. If a node which receives NR(0,1) is in fact not indicating a local error, the receive node can take the received NR(0,1) as an indication that there is no error in the protection domain, and recovery procedures (WTR or DNR) should begin.

This is addressed by adding the following text as the penultimate bullet in section 4.3.3.4:

- o If a node is in Protecting Failure state due to a remote SF-W and receives NR(0,1), this SHALL cause the node to begin recovery procedures. If the LER is configured for revertive behavior, it enters into Wait-to-Restore state, starts the WTR timer, and begins transmitting WTR(0,1). If the LER is configured for non-revertive behavior, it enters into Do-Not-Revert state and begins transmitting a DNR(0,1) message.

Additionally, the final bullet in section 4.3.3.3 is changed from

- o A remote NR(0,0) message SHALL be ignored if in local Protecting administrative state.

to

- o A remote No Request message SHALL be ignored if in local Protecting administrative state.

This indicates that a remote NR triggers the same behavior regardless of the value of FPath and Path. This change does not directly address issue #8, but fixes a similar issue - if a node receives NR while in Remote administrative state, the value of FPath and Path have no bearing on the node's reaction to this NR.

4. Clarifying PSC's behavior in the face of multiple inputs

RFC6378 describes the PSC state machine. Figure 1 in section 3 shows two inputs into the PSC Control logic - Local Request logic and Remote PSC Request. When there is only one input into the PSC Control logic - a local request or a remote request but not both - the PSC Control logic decides what that input signifies and then takes one or more actions, as necessary. This is what the PSC State Machine in section 4.3 describes.

RFC6378 does not sufficiently describe the behavior in the face of multiple inputs into the PSC Control Logic (one Local Request and one

Remote Request). This section clarifies the desired behavior.

There are two cases to think about when considering dual inputs into the PSC Control logic. The first is when the same request is presented from both local and remote sources. One example of this case is a failure of the Working LSP. A bidirectional fiber cut will result in the PSC Control logic receiving both a local SF-W (due to loss of light on the underlying fiber) and a remote SF-W (due to the peer node's loss of light). Incidentally, a unidirectional fiber cut will very likely result in a bidirectional failure scenario as it is expected that most MPLS-TP deployments will be running MPLS OAM [RFC6428]. For convenience, this scenario is written as [L(FS), R(FS)]

The second case, which is handled in exactly the same way as the first, is when the two inputs into the PSC Control logic describe different events. There are a number of variations on this case. One example is when there is a Forced Switch (that is, a forced switch to the protection LSP) coming from the Local request logic and a Lockout of Protection from the Remote PSC Request. This is shortened to [L(FS), R(LO)].

In both cases the question is not how the PSC Control logic decides which of these is the one it acts upon. Section 4.3.2 of RFC6378 lists the priority order, and prioritizes the local input over the remote input in case both inputs are of the same priority. So in the first example it is the local SF that drives the PSC Control logic, and in the second example it is the Lockout which drives the PSC Control logic.

The point that this section clears up is around what happens when the highest priority input goes away. Consider the first case. Initially, the PSC Control logic has [L(FS), R(FS)] and R(FS) is driving PSC's behavior. When L(FS) is removed but R(FS) remains, what does PSC do? A strict reading of the FSM would suggest that PSC transition from PA:F:L into N, and at some future time (perhaps after the remote request refreshes) PSC would transition from N to PA:F:R. This is clearly an unreasonable behavior, as there is no sensible justification for a node behaving as if things were normal (i.e. N state) when it is clear that they are not.

The second case is similar. If a node starts with [L(LO), R(FS)] and the local lockout is removed, a strict reading of the state machine would suggest that the node transition from UA:LO:L to N, and then at some future time presumably notice the R(FS) and transition from N to PA:F:R. As with the first case, this is clearly not a useful behavior.

In both cases, the request which was driving PSC's behavior was removed. What should happen is that the PSC Control logic should, upon removal of an input, reevaluate all other inputs to decide on the next course of action.

There is a third case. Consider a node with [L(FS), R(LO)]. At some point in time the remote node replaces its Lockout request with a Signal Fail on Working, so that the inputs into the PSC Control logic on the receiving node go immediately to [L(FS), R(SF-W)]. Similar to the first two cases, the node should reevaluate both its local and remote inputs to determine the highest priority among them, and act on that input accordingly. That is in fact what happens, as defined in Section 4.3.3:

"When a LER is in a remote state, i.e., state transition in reaction to a PSC message received from the far-end LER, and receives a new PSC message from the far-end LER that indicates a contradictory state, e.g., in remote Unavailable state receiving a remote FS(1,1) message, then the PSC Control logic SHALL reevaluate all inputs (both the local input and the remote message) as if the LER is in the Normal state."

This section amends that paragraph to handle the first two cases. The essence of the quoted paragraph is that when faced with multiple inputs, PSC must reevaluate any changes as if it was in Normal state. So the quoted paragraph is replaced with the following text:

"The PSC Control logic may simultaneously have Local and Remote requests, and the highest priority of these requests ultimately drives the behavior of the PSC Control logic. When this highest priority request is removed or is replaced with another input, then the PSC Control logic SHALL reevaluate all inputs (both the local input and the remote message), transitioning into a new state only upon reevaluation of all inputs".

5. Security Considerations

These changes and clarifications raise no new security concerns.

6. IANA Considerations

None.

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

7. Acknowledgements

The author of this document thanks Annamaria Fulignoli, Sagar Soni, George Swallow and Yaacov Weingarten for their contributions and review.

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.
- [RFC6378] Weingarten, Y., Bryant, S., Osborne, E., Sprecher, N., and A. Fulignoli, "MPLS Transport Profile (MPLS-TP) Linear Protection", RFC 6378, October 2011.
- [RFC6428] Allan, D., Swallow Ed. , G., and J. Drake Ed. , "Proactive Connectivity Verification, Continuity Check, and Remote Defect Indication for the MPLS Transport Profile", RFC 6428, November 2011.

8.2. Informative References

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

Author's Address

Eric Osborne

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: August 20, 2013

L. Andersson
M. Chen
Huawei
T. Petch
Engineering Networks Ltd
February 16, 2013

"MPLS LSP Ping TLVs and sub-TLVs registry"
draft-pac-mpls-lsp-ping-tlvs-and-sub-tlvs-registry-00.txt

Abstract

This document addresses issues with the structure, allocation policies and clarity in the use of the "TLVs and sub-TLVs" of the "Multi-Protocol Label Switching (MPLS) Label Switched Paths (LSPs) Ping Parameters" in the "Multiprotocol Label Switching Architecture (MPLS)" name space.

This document does not change any existing allocations and the new structure is backwards compatible with the existing registries.

The policy for the allocation of TLVs is unchanged but future allocations of sub-TLVs will come from a single namespace, common to all TLVs of LSP Ping Parameters.

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 August 20, 2013.

Copyright Notice

Copyright (c) 2013 IETF Trust and the persons identified as the document authors. All rights reserved.

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

Table of Contents

| | |
|---|----|
| 1. Introduction | 4 |
| 2. Current situation | 5 |
| 2.1. Current situation - model | 5 |
| 2.2. Allocation policies and Scope | 6 |
| 2.3. A closer look at the model | 6 |
| 3. New registry structure | 9 |
| 3.1. If we'd done it from start | 9 |
| 3.2. TLV Registry and allocation procedures | 10 |
| 3.3. Sub-TLV registries and allocation policies | 12 |
| 3.3.1. Sub-TLV registry for all TLVs | 12 |
| 3.3.2. Sub TLV registry for TLV Type 9 | 14 |
| 3.3.3. Sub TLV registry for TLV Type 11 | 15 |
| 3.3.4. Sub TLV registry for TLV Type 20 | 16 |
| 4. Security Considerations | 17 |
| 5. IANA considerations | 18 |
| 6. Acknowledgments | 20 |
| 7. References | 21 |
| 7.1. Normative References | 21 |
| 7.2. Informative references | 21 |
| Authors' Addresses | 22 |

1. Introduction

This document revises the allocation policies in the use of the TLVs and sub-TLVs of the MPLS LSP Ping Parameters, as defined in [RFC4379].

This document does not change any existing allocations and the new structure is backwards compatible with the existing registries.

The policy for the allocation of TLVs is unchanged but future allocations of sub-TLVs will come from a single namespace, common to all TLVs of MPLS LSP Ping Parameters.

The allocation of existing sub-TLVs is unaltered, so that the meaning of, e.g., sub-TLV sub-Type 1 is dependent on the TLV under which it appears. No future allocations will be made with a sub-Type of less than 32. Future allocations will be made from a single namespace starting at 32; a sub-TLV defined in this way may appear as part of any current or future TLV. The document that specifies such an allocation should state which TLVs the sub-TLV may appear under and indicate any other future use which seems appropriate or inappropriate.

2. Current situation

Today all TLVs and sub-TLVs are found in a single table, and the allocation policies are the same for all TLVs and sub-TLVs. The table below illustrates how the registry is set up.

Initially this might have been a good idea, but over time, with an increasing number of TLVs, and with some sub-TLVs shared across TLVs, it has become increasingly difficult to understand how the allocation policies interact.

2.1. Current situation - model

The table below illustrates how the registry is set up and the allocation policies work currently. We have chosen not to just copy the current registry here, but instead build a model that shows how the allocation policies work.

--Note to RFC Editor; the various RFC aaaa to RFC zzzz are really meant to be like that in the finished document; we are not asking you to replace them with anything:-)

Current TLV and sub-TLV registry (model)

| Type | Sub-type | Value field | Reference |
|------|----------|-------------------------|---------------|
| 1 | | TLV # 1 | RFC xxxx (1) |
| 1 | 1 | sub-TLV # 1 | RFC xxxx (2) |
| 1 | 2 | sub-TLV # 2 | RFC yyyy (3) |
| 1 | 3 | sub-TLV # 3 | RFC yyyy (4) |
| 2 | | TLV # 2 | RFC xxxx (5) |
| 3 | | TLV # 3 | RFC zzzz (6) |
| 3 | 1 | sub-TLV # 1 | RFC zzzz (7) |
| 3 | 2 | sub-TLV # 2 | RFC zzzz (8) |
| 3 | 3 | sub-TLV # 3 | RFC aaaa (9) |
| 4 | | TLV # 4 | RFC bbbb (10) |
| 4 | 1-16383 | as specified for type 1 | RFC bbbb (11) |
| 5 | | TLV # 5 | RFC cccc (12) |
| 5 | 1-65535 | as specified for type 1 | RFC cccc (13) |

Note: The row number column to the right is added here to discuss what is on the different rows.

2.2. Allocation policies and Scope

TLV and sub-TLV registration procedures

| Range | Registration Procedures | Notes |
|-------------|-------------------------|--|
| 0-16383 | Standards Action | This range is for mandatory TLVs or for optional TLVs that require an error message if not recognized. |
| 16384-31743 | Specification Required | Experimental RFC needed |
| 31744-32767 | Vendor Private Use | MUST NOT be allocated |
| 32768-49161 | Standards Action | This range is for optional TLVs that can be silently dropped if not recognized. |
| 49162-64511 | Specification Required | Experimental RFC needed |
| 64512-65535 | Vendor Private Use | MUST NOT be allocated |

The IANA registry does not give enough information to correctly allocate TLVs and sub-TLVs, instead careful reading of [RFC4379] is necessary.

[RFC4379] says:

The valid range for TLVs and sub-TLVs is 0-65535. Assignments in the range 0-16383 and 32768-49161 are made via Standards Action as defined in Section 5; assignments in the range 16384-31743 and 49162-64511 are made via "Specification Required" as defined above; values in the range 31744-32767 and 64512-65535 are for Vendor Private Use, and MUST NOT be allocated.

[RFC4379] also says that the sub-TLVs are scoped by the TLVs, i.e. a sub-TLV defined for one TLV is valid for that TLV only. Later the practice to re-define (a block of) sub-TLVs defined for one TLV for another TLV was introduced.

2.3. A closer look at the model

The list below contains what we see as the results of the most common allocation requests for this registry.

1. Row 1 says that IANA has allocated a TLV as requested in RFCxxxx. This TLV is type 1.

RFCxxxx is the document that defines the registry and sets up the allocation policies.
2. Row 2 says that IANA has allocated a sub-TLV for TLV type 1, "sub-TLV #1", the source for this allocation is the same that defined the registry and allocated the TLV Type 1 (RFCxxxx).
3. Row 3 says that IANA has allocated a second sub-TLV (sub-TLV #2) for TLV type 1, the source for this allocation is RFCyyyy.

-

4. Row 4 says that IANA has allocated a third sub-TLV (sub-TLV #3) for TLV type 1, the source for this allocation is RFCyyyy.

-

5. Row 5 says that IANA has allocated a new TLV (TLV type 2), the source for this allocation is RFCxxxx, the same RFC that defined the registry.

TLV type 2 has no sub-TLVs yet defined.

6. Row 6 says that IANA has allocated a new TLV (TLV type 3), the source for this allocation is RFCzzzz.

-

7. Row 7 says that IANA has allocated a sub-TLV (sub-TLV # 1) for TLV type 3, the source for this allocation is RFCzzzz.

This means that we have one sub-TLV # 1 for TLV type 1, and another sub-TLV # 1 for TLV type 3. In itself this is not a problem, the sub-TLVs are scoped by the TLVs.

8. Row 8 says that IANA has allocated a sub-TLV (sub-TLV # 2) for TLV type 3, the source for this allocation is RFCzzzz.

-

9. Row 9 says that IANA has allocated a sub-TLV (sub-TLV # 2) for TLV type 3, the source for this allocation is RFCaaaa.

-

10. Row 10 says that IANA has allocated a new TLV (TLV type 4), the source for this allocation is RFCbbbb.

-

11. Row 11 says that IANA has been instructed not to allocate any sub-TLVs from the range 1-16383, but that the sub-TLVs for TLV type 4, shall use the same sub-TLVs that have been specified for TLV type 1 in this range.

This implies that other ranges for TLV type 4 are open for allocation for "TLV type 4 specific sub-TLVs". This is specified in RFCbbbb.

12. Row 12 says that IANA has allocated a new TLV (TLV type 5), the source for this allocation is RFCcccc.

-

13. Row 13 says that IANA has been instructed not to allocate any sub-TLVs from the entire range (1-65535), but that the sub-TLVs for TLV type 5, shall use the same sub-TLVs that have been specified for TLV type 1. This is specified in RFCcccc.

Close reading of the allocation rules would likely show that disallowing the assignment of vendor-specific sub-TLVs is moot.

3. New registry structure

3.1. If we'd done it from start

The name space of sub-TLVs is very large, 65 535 potential TLVs times 65 535 sub-TLVs per TLV, gives a maximum of 4 294 836 335 sub-TLVs.

There seems no reason why that number of sub-TLVs should be needed; rather, 65 535 sub-TLVs shared among all TLVs would seem to have been more than sufficient. If the IANA registries had been set up with one registry for TLVs and another for sub-TLVs, that would have resulted in registries and allocation policies much easier to understand and comprehend.

In practice, the same sub-TLV number appears more than once under different TLVs with a different meaning on each occasion. Thus sub-TLV 1 appears under TLV Type 1 as LDP IPv4 Prefix, under TLV Type 11 as IPv4 Egress Address P2MP Responder and under TLV Type 20 as Multipath data. At the same time, TLVs Types 16 and 21 reuse sub-TLV 1 with the same meaning as for TLV Type 1.

Thus it is now impossible to create a single registry for sub-TLVs which encompasses all existing sub-TLVs. At the same time, such a registry would simplify future registration and use, allowing, for example, a sub-TLV to be defined for an IPv6 address that would then be used wherever such an address is required. Hence, the future policy for the registration of sub-TLVs is to have a single registry regardless of which TLV the sub-TLV appears under. This registry follows the same pattern as the existing registries, namely of

| | | |
|-------------|------------------------|--------------------------------------|
| 0-16383 | Standards Action | Mandatory (sub)TLVs |
| 16384-31743 | Specification Required | Mandatory Experimental
RFC needed |
| 31744-32767 | Vendor Private Use | MUST NOT be allocated |
| 32768-49161 | Standards Action | Optional (sub)TLVs |
| 49162-64511 | Specification Required | Optional Experimental
RFC needed |
| 64512-65535 | Vendor Private Use | MUST NOT be allocated |

excepting that the range 0 to 31 is now reserved and MUST NOT be assigned lest there is an overlap with existing definitions. The

choice of 32 is somewhat greater than the greatest, existing, defined sub-TLV, 25 for TLV Type 1, and is chosen to be a more user-friendly, easier to remember, number than, say, 26 or 29.

The examples in TLV Registry and allocation procedures (Section 3.2) and Sub-TLV registries and allocation policies (Section 3.3) are the actual allocations in the IANA registry as they are found at the time of writing of this document (January 2013).

3.2. TLV Registry and allocation procedures

TLV registration procedures

| Range | Registration Procedures | Notes |
|-------------|-------------------------|---|
| 0-16383 | Standards Action | This range is for mandatory TLVs or for optional TLVs that require an error message if not recognized. |
| 16384-31743 | Specification Required | Experimental RFC needed
This range is for mandatory TLVs or for optional TLVs that require an error message if not recognized. |
| 31744-32767 | Vendor Private Use | MUST NOT be allocated |
| 32768-49161 | Standards Action | This range is for optional TLVs that can be silently discarded if not recognized. |
| 49162-64511 | Specification Required | Experimental RFC needed
This range is for optional TLVs that can be silently discarded if not recognized. |
| 64512-65535 | Vendor Private Use | MUST NOT be allocated |

LSP Ping TLV Registry

| Type | Value Field | Reference |
|------|------------------|-----------|
| 1 | Target FEC Stack | [RFC4379] |

| | | |
|-------------|------------------------------------|------------------------|
| 2 | Downstream Mapping
(DEPRECATED) | [RFC4379]
[RFC6424] |
| 3 | Pad | [RFC4379] |
| 4 | Not Assigned | [RFC4379] |
| 5 | Vendor Enterprise Number | [RFC4379] |
| 6 | Not Assigned | [RFC4379] |
| 7 | Interface and Label Stack | [RFC4379] |
| 8 | Not Assigned | [RFC4379] |
| 9 | Errored TLVs | [RFC4379] |
| 10 | Reply TOS Byte | [RFC4379] |
| 11 | P2MP Responder Identifier | [RFC6425] |
| 12 | Echo Jitter | [RFC6425] |
| 13 | Source ID | [RFC6426] |
| 14 | Destination ID | [RFC6426] |
| 15 | BFD Discriminator | [RFC5884] |
| 16 | Reverse-path Target FEC Stack | [RFC6426] |
| 17-19 | Unassigned | |
| 20 | Downstream Detailed Mapping | [RFC6424] |
| 22-31743 | Unassigned | |
| 31744-32767 | Reserved for Vendor
private use | [RFC4379] |
| 32768-64511 | Unassigned | |
| 64512-65535 | Reserved for Vendor
private use | [RFC4379] |

3.3. Sub-TLV registries and allocation policies

3.3.1. Sub-TLV registry for all TLVs

Registration procedures for all sub-TLVs

| Range | Registration Procedures | Notes |
|-------------|-------------------------|---|
| 0-31 | Reserved | Existing allocations in this range are unaltered.
No future allocations are to be made from this range. |
| 32-16383 | Standards Action | This range is for mandatory sub-TLVs or for optional sub-TLVs that require an error message if not recognized. |
| 16384-31743 | Specification Required | Experimental RFC needed
This range is for mandatory sub-TLVs or for optional sub-TLVs that require an error message if not recognized. |
| 31744-32767 | Vendor Private Use | MUST NOT be allocated |
| 32768-49161 | Standards Action | This range is for optional sub-TLVs that can be silently discarded if not recognized. |
| 49162-64511 | Specification Required | Experimental RFC needed
This range is for optional sub-TLVs that can be silently discarded if not recognized. |
| 64512-65535 | Vendor Private Use | MUST NOT be allocated |

Type 1 TLV sub-TLVs

Sub-TLVs for TLV Type 1

| Sub-TLV | Value Field | Reference |
|---------|--------------------------|---------------|
| 0 | Reserved - do not assign | This document |

| | | |
|----|---|------------------------|
| 1 | LDP IPv4 prefix | [RFC4379] |
| 2 | LDP IPv6 prefix | [RFC4379] |
| 3 | RSVP IPv4 LSP | [RFC4379] |
| 4 | RSVP IPv6 LSP | [RFC4379] |
| 5 | Not Assigned | [RFC4379] |
| 6 | VPN IPv4 prefix | [RFC4379] |
| 7 | VPN IPv6 prefix | [RFC4379] |
| 8 | L2 VPN endpoint | [RFC4379] |
| 9 | "FEC 128" Pseudowire - IPv4
(DEPRECATED) | [RFC4379]
[RFC6829] |
| 10 | "FEC 128" Pseudowire - IPv4 | [RFC4379]
[RFC6829] |
| 11 | "FEC 129" Pseudowire - IPv4 | [RFC4379]
[RFC6829] |
| 12 | BGP labeled IPv4 prefix | [RFC4379] |
| 13 | BGP labeled IPv6 prefix | [RFC4379] |
| 14 | Generic IPv4 prefix | [RFC4379] |
| 15 | Generic IPv6 prefix | [RFC4379] |
| 16 | Nil FEC | [RFC4379] |
| 17 | RSVP P2MP IPv4 Session | [RFC6425] |
| 18 | RSVP P2MP IPv6 Session | [RFC6425] |
| 19 | Multicast P2MP LDP FEC Stack | [RFC6425] |
| 20 | Multicast MP2MP LDP FEC Stack | [RFC6425] |
| 21 | Unassigned | |
| 22 | Static LSP | [RFC6426] |

| | | |
|----|-----------------------------|-----------|
| 23 | Static Pseudowire | [RFC6426] |
| 24 | "FEC 128" Pseudowire - IPv6 | [RFC6829] |
| 25 | "FEC 129" Pseudowire - IPv6 | [RFC6829] |

3.3.2. Sub TLV registry for TLV Type 9

TLV Type 9 has a very different allocation policy to all other TLVs; any value carried in the Value field of the TLV is a copy of a TLV that has not been understood or recognized. It is even doubtful that "All values" technically is a sub-TLV, but both the IANA registry and [RFC4379] says it is. Equally, it is unclear whether or not TLV Type 9 should be used to report a sub-TLV that has not been recognised and if it is, how that sub-TLV should appear in the Type 9 TLV. More work on this is needed but that falls outside the scope of this document.

Registration procedures TLV type 9 sub-TLVs

| Range | Registration Procedures | Notes |
|---------|-------------------------------|--|
| 0-65535 | Reserved MUST NOT be assigned | Any value carried in the value field of TLV type 9 means that a TLV has not been understood. |

Type 9 TLV sub-TLVs

Sub-TLVs for TLV Type 9

| Sub-TLV | Value Field | Reference |
|------------|----------------------------|-----------|
| All values | TLV that is not understood | [RFC4379] |

3.3.3. Sub TLV registry for TLV Type 11

Registration procedures TLV type 11 sub-TLVs
(as specified by RFC6425)

| Range | Registration Procedures | Notes |
|-------------|-------------------------|--|
| 0-16383 | Standards Action | This range is for mandatory TLVs or for optional TLVs that require an error message if not recognized. |
| 16384-31743 | Specification Required | Experimental RFC needed |
| 31744-32767 | Vendor Private Use | MUST NOT be allocated |
| 32768-49161 | Standards Action | This range is for optional TLVs that can be silently dropped if not recognized. |
| 49162-64511 | Specification Required | Experimental RFC needed |
| 64512-65535 | Vendor Private Use | MUST NOT be allocated |

Type 11 TLV sub-TLVs

| sub-TLV | Value Field | Reference |
|---------|------------------------------------|---------------|
| 0 | Reserved not to be assigned | This document |
| 1 | IPv4 Egress Address P2MP Responder | [RFC6425] |
| 2 | IPv6 Egress Address P2MP Responder | [RFC6425] |
| 3 | IPv4 Node Address P2MP Responder | [RFC6425] |
| 4 | IPv6 Node Address P2MP | [RFC6425] |

| | | |
|-------|-----------|-------|
| | Responder | |
| ----- | ----- | ----- |

3.3.4. Sub TLV registry for TLV Type 20

Registration procedures TLV type 20 sub-TLVs
(as specified by RFC6424)

| Range | Registration Procedures | Notes |
|-------------|-------------------------|--|
| 0-16383 | Standards Action | This range is for mandatory TLVs or for optional TLVs that require an error message if not recognized. |
| 16384-31743 | Specification Required | Experimental RFC needed |
| 31744-32767 | Vendor Private Use | MUST NOT be allocated |
| 32768-49161 | Standards Action | This range is for optional TLVs that can be silently dropped if not recognized. |
| 49162-64511 | Specification Required | Experimental RFC needed |
| 64512-65535 | Vendor Private Use | MUST NOT be allocated |

Type 20 TLV sub-TLVs

| sub-TLV | Value Field | Reference |
|---------|------------------|-----------|
| 1 | Multipath data | [RFC6424] |
| 2 | Label stack | [RFC6424] |
| 3 | FEC stack change | [RFC6424] |

4. Security Considerations

This document amends the policy for the registration of sub-TLVs of MPLS LSP Ping. As such, it does not introduce any additional security considerations over and above those included with the specification of the sub-TLVs themselves.

5. IANA considerations

This document revises the allocation policies in the use of the TLVs and sub-TLVs of the MPLS LSP Ping Parameters, as previously defined in [RFC4379].

The allocation policy for TLVs is unaltered from RFC4379 but the IANA registry should be updated to refer to this document, lest users of this information do not appreciate that the policies for sub-TLVs, as specified in [RFC4379], no longer apply; that is, users are directed here first, so that they have the current, overall procedures.

The allocation policy for sub-TLVs is that all sub-TLVs now come from a common pool so that a sub-TLV sub-Type number is now unique within all of MPLS LSP Ping Parameters.

The lowest value for allocation of any sub-TLV sub-Type is 32, so as to avoid overlap with any sub-TLV Type currently defined or under consideration.

The registration procedure is as specified in Sub-TLV registry for all TLVs (Section 3.3.1), namely

| Range | Registration Procedures | Notes |
|-------------|-------------------------|---|
| 0-31 | Reserved | Existing allocations in this range are unaltered.
No future allocations are to be made from this range. |
| 32-16383 | Standards Action | This range is for mandatory sub-TLVs or for optional sub-TLVs that require an error message if not recognized. |
| 16384-31743 | Specification Required | Experimental RFC needed
This range is for mandatory sub-TLVs or for optional sub-TLVs that require an error message if not recognized. |
| 31744-32767 | Vendor Private Use | MUST NOT be allocated |
| 32768-49161 | Standards Action | This range is for optional sub-TLVs that can be silently discarded if not recognized. |
| 49162-64511 | Specification Required | Experimental RFC needed
This range is for optional sub-TLVs that can be silently discarded if not recognized. |
| 64512-65535 | Vendor Private Use | MUST NOT be allocated |

6. Acknowledgments

TBD

7. References

7.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC4379] Kompella, K. and G. Swallow, "Detecting Multi-Protocol Label Switched (MPLS) Data Plane Failures", RFC 4379, February 2006.

7.2. Informative references

- [RFC5884] Aggarwal, R., Kompella, K., Nadeau, T., and G. Swallow, "Bidirectional Forwarding Detection (BFD) for MPLS Label Switched Paths (LSPs)", RFC 5884, June 2010.
- [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.
- [RFC6426] Gray, E., Bahadur, N., Boutros, S., and R. Aggarwal, "MPLS On-Demand Connectivity Verification and Route Tracing", RFC 6426, November 2011.
- [RFC6829] Chen, M., Pan, P., Pignataro, C., and R. Asati, "Label Switched Path (LSP) Ping for Pseudowire Forwarding Equivalence Classes (FECs) Advertised over IPv6", RFC 6829, January 2013.

Authors' Addresses

Loa Andersson
Huawei

Email: loa@mail01.huawei.com

Mach(Guoyi) Chen
Huawei

Email: mach.chen@huawei.com

Tom Petch
Engineering Networks Ltd

Email: tomSecurity@network-engineer.co.uk

Network Working Group
Internet-Draft
Updates: 4928, 6790 (if approved)
Intended status: Standards Track
Expires: August 20, 2013

C. Pignataro
Cisco Systems, Inc.
L. Andersson
Huawei Technologies
K. Kompella
Juniper Networks
February 16, 2013

The Use of MPLS Special Purpose Labels for the Computation of Load
Balancing
draft-pignataro-mpls-reserved-labels-lb-01

Abstract

In addition to being used for forwarding, an MPLS label stack may also be used as an entropy source to perform load balancing computation in various ways. RFC 4928 and RFC 6790 describe this mechanism in great detail. However, those two RFCs differ in the use of MPLS special purpose labels (previously referred to as "reserved labels") for computation of load balancing. This document addresses this difference in specifications by providing a more comprehensive set of recommendations.

This document updates RFC 4928 and RFC 6790.

Requirements Language

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

Status of this Memo

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

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

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

This Internet-Draft will expire on August 20, 2013.

Copyright Notice

Copyright (c) 2013 IETF Trust and the persons identified as the document authors. All rights reserved.

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

Table of Contents

| | |
|---|---|
| 1. Introduction | 3 |
| 2. MPLS Special Purpose Labels and Load Balancing | 3 |
| 2.1. Current Specifications | 3 |
| 2.2. Detail of Updates | 4 |
| 3. IANA Considerations | 4 |
| 4. Security Considerations | 4 |
| 5. Acknowledgements | 4 |
| 6. Normative References | 5 |
| Authors' Addresses | 5 |

1. Introduction

In addition to being used for forwarding, an MPLS label stack may also be used as an entropy source to perform load balancing computation in various ways. RFC 4928 [RFC4928] and RFC 6790 [RFC6790] describe this mechanism in great detail. However, those two RFCs differ in the use of MPLS special purpose labels (previously referred to as "reserved labels") for computation of load balancing. This document addresses this difference in specifications by providing a more comprehensive set of recommendations.

This document updates RFC 4928 and RFC 6790.

2. MPLS Special Purpose Labels and Load Balancing

2.1. Current Specifications

This section highlights current specifications relating to the usage of MPLS special purpose labels for purposes of load balancing computation.

[RFC4928] states that special purpose labels ("reserved labels") may be used for load balancing, and describes current ECMP practice as follows:

It must also be noted that LSRs that correctly identify a payload as not being IP most often will load-share traffic across multiple equal-cost paths based on the label stack. Any reserved label, no matter where it is located in the stack, may be included in the computation for load balancing. Modification of the label stack between packets of a single flow could result in re-ordering that flow. That is, were an explicit null or a router-alert label to be added to a packet, that packet could take a different path through the network.

[RFC6790], conversely, succinctly states that special purpose labels ("reserved labels") MUST NOT be used for load balancing:

If a transit LSR recognizes the ELI, it MAY choose to load balance solely on the following label (the EL); otherwise, it SHOULD use as much of the whole label stack as feasible as keys for the load-balancing function. In any case, reserved labels MUST NOT be used as keys for the load-balancing function.

2.2. Detail of Updates

There are several MPLS special purpose labels. MPLS special purpose labels have special meaning both in the control plane and the data plane, including an indication for OAM. OAM packets not taking the same path as data packets defeats their purpose.

On the other hand, it is existing practice that MPLS equipment load balances on the full label stack, or on portions of the full label stack irrespective of the value of the label, as documented in [RFC4928]. A new specification cannot automatically render obsolete equipment that conformed to a prior documented specification.

Consequently, this document updates RFC 4928 and RFC 6790 by specifying that:

1. It is RECOMMENDED that new implementations of MPLS equipment do not use MPLS special purpose labels as input into the load balancing computation.
2. MPLS forwarding equipment SHOULD document their load-balancing behavior in presence of MPLS special purpose labels.

3. IANA Considerations

This document makes no request of IANA.

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

4. Security Considerations

This document updates RFC 4928 and RFC 6790 by providing a more comprehensive set of recommendation on the use of MPLS special purpose labels as input into the load-balancing computations. The security considerations of these two RFCs are unchanged. This update does not impose any new security considerations.

5. Acknowledgements

The authors would like to thank thorough reviews and useful comments and suggestions from Stewart Bryant, Adrian Farrel, and John E. Drake.

6. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC4928] Swallow, G., Bryant, S., and L. Andersson, "Avoiding Equal Cost Multipath Treatment in MPLS Networks", BCP 128, RFC 4928, June 2007.
- [RFC6790] Kompella, K., Drake, J., Amante, S., Henderickx, W., and L. Yong, "The Use of Entropy Labels in MPLS Forwarding", RFC 6790, November 2012.

Authors' Addresses

Carlos Pignataro
Cisco Systems, Inc.

Email: cpignata@cisco.com

Loa Andersson
Huawei Technologies

Email: loa@mail01.huawei.com

Kireeti Kompella
Juniper Networks

Email: kireeti.kompella@gmail.com

MPLS Working Group
INTERNET-DRAFT
Intended Status: Proposed Standard
Expires: August 22, 2013

R. Singh
Y. Shen
J. Drake
Juniper Networks
February 18, 2013

Entropy label for seamless MPLS
draft-ravisingh-mpls-el-for-seamless-mpls-00

Abstract

This document describes how entropy labels can be used for load balancing in a seamless MPLS architecture. The definition of the control plane and data plane behavior at LSP stitching points; and at the ingress of an LSP in a hierarchy of LSPs, as described in this document, brings the benefits of entropy labels to seamless MPLS as MPLS deployments proliferate in the access and aggregation networks.

This document updates RFC 6790.

Status of this Memo

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

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

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

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

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

This Internet-Draft will expire on August 22, 2013.

Copyright and License Notice

Copyright (c) 2013 IETF Trust and the persons identified as the document authors. All rights reserved.

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

Table of Contents

| | | |
|---------|---|----|
| 1 | Introduction | 4 |
| 2 | Terminology | 5 |
| 3 | Key attributes of the entropy label solution: Summary from [EL-RFC] | 6 |
| 4 | Problems and Motivation | 6 |
| 4.1 | EL applicability for seamless MPLS | 7 |
| 4.2 | EL for LSP stitching | 7 |
| 4.2.1 | Spectrum of EL usage behaviors required to be supported for stitched LSPs | 8 |
| 4.2.1.1 | Entropy label for per-segment LSP | 9 |
| 4.2.1.2 | Entropy label for notional-segment-LSP(s) | 9 |
| 4.2.1.3 | Entropy label for e2e LSP | 10 |
| 4.3 | EL for LSP hierarchy | 10 |
| 4.3.1 | Possibility of unnecessary reduction of max-payload of the LSP: | 10 |
| 4.3.2 | Possibility of EL being non-usable for load-balancing: | 11 |
| 5 | EL for LSP stitching/hierarchy | 13 |
| 5.1 | Additional EL abstractions: specific to LSP stitching/hierarchy | 13 |
| 5.2 | New abstractions: EL applicability for LSP stitching | 13 |
| 5.2.1 | Signaling | 13 |
| 5.2.1.1 | Signaling ELC at stitching points (Translation rules) | 14 |
| 5.2.2 | Data plane aspects | 15 |
| 5.2.2.1 | Stitching: Differing EL dispositions | 15 |
| 5.3 | New abstractions: EL applicability for LSP hierarchy | 18 |
| 5.3.1 | Management plane: | 18 |
| 5.3.2 | Data plane aspects | 18 |
| 6 | Security considerations | 19 |
| 7 | Acknowledgments | 19 |
| 8 | IANA considerations | 19 |

| | | |
|-----|----------------------------------|----|
| 9 | References | 19 |
| 9.1 | Normative References | 19 |
| 9.2 | Informative References | 20 |
| | Authors' Addresses | 21 |

1 Introduction

[EL-RFC] specifies a way to implement load-balancing in an MPLS network such that sub-flows of an LSP may be identified and sent on different paths through the network. This is achieved by using entropy labels (ELs) to abstract out the flow-identifying information into the entropy label and inserting the entropy label underneath the LSP label. The transit LSRs perform the load-balancing hash-computation, on the label-stack alone, to effect a good load-balancing outcome without a need to parse inner headers.

The key feature of [EL-RFC] is that it defines the EL in the context of a given LSP. [EL-RFC] defines the signaling extensions by which entropy label capability might be signaled for LSPs setup by RSVP-TE, LDP or [LU-BGP]. While that works well for individual LSPs, there are additional issues to consider for the seamless MPLS architecture [S-MPLS].

The currently-under-definition framework for seamless MPLS proposes an architecture ([S-MPLS]) that shall enable the setting-up of MPLS LSPs from access nodes to access nodes using a medley of signaling protocols and statically configured LSPs. There are special EL-related considerations that need to be dealt with to make EL more suitable for seamless MPLS.

This document defines additional abstractions and rules for the use of entropy-label with LSP stitching/hierarchy to enable the use of ELs for the seamless MPLS architecture. This document describes how entropy labels may be used when the LSP has been setup by stitching LSP segments or by tunneling the LSP over other LSPs. It is conceivable that different signaling protocols are in use to create an e2e LSP.

LSP stitching is the process of connecting LSP segments in the data plane to form a single e2e data plane LSP. This is achieved by setting up LSP segments through signaling or through management action, and then signaling an e2e LSP that "uses" these LSP segments as hops in its path. The procedures for LSP stitching are described in [STITCHING]. Labeled data traffic flowing over e2e MPLS LSPs, that have been setup using multiple different protocols (by stitching together segments), would benefit from having the entropy label be included in it.

LSP hierarchy is defined in [MPLS-ARCH] and [GMPLS-HIER]. Usage of entropy label in LSP hierarchies has some peculiar practical issues that will benefit from some additional flexibility in inserting ELs for a specific layer in an LSP hierarchy.

2. Terminology

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

The following acronyms/terms are used:

e2e: End to end LSP that has been setup by stitching together LSP segments

ECMP: Equal Cost Multi-Path

EL: Entropy Label

ELC: Entropy Label Capability or Entropy Label Capable

ELI: Entropy Label Indicator

Intrinsic ELC: Entropy label capability/capable as in [EL-RFC]. In this document, an LSP is considered to be "intrinsically" EL-capable when the:

- the ingress of the LSP has the ability to compute and PUSH the EL before PUSHing the ELI before PUSHing the LSP label; and
- the egress/PHR of the LSP-segment has the ability to POP the (ELI+EL) at the egress/PHR while POPping-transport-label/ELI-is-top-label respectively.

LAG: Link Aggregation Group

LER: Label Edge Router

LSP: Label Switched Path

LSR: Label Switching Router

Notional ingress: Ingress LER for an LSP segment that is inserting the (ELI+EL) on data traffic going over an e2e LSP

Notional egress: egress LER for an LSP segment that is removing the (ELI+EL) from data traffic going over an e2e LSP

Notional LSP segment: the portion of the e2e LSP between a notional ingress and a notional egress

PHP: Penultimate Hop Popping

PHR: Penultimate Hop Router

UHP: Ultimate Hop Popping

NOTE: this document references the (ELI+EL) pair simply as EL when the presence of the ELI is of no significance for the issue being described. The presence of ELI is mandatory as per [EL-RFC] when EL is in use.

3. Key attributes of the entropy label solution: Summary from [EL-RFC]

- Transport-label-PUSHing router inserts (ELI+EL)
The (ELI+EL) insertion is done, if at all, by a router that is PUSHing the transport LSP's label.
- Ingress-LER (transport-label-PUSHing-router) inserts (ELI+EL) only if the PHR/egress has signaled ability to strip it off.
- Transport-label-POPing router POPs (ELI+EL) PHR/egress of the LSP is responsible for POPing off the (ELI+EL) after it has been exposed as the top label on the packet due to POPing the transport label. The removal of the (ELI+EL) is done either when the ELI is the top label; or when the ELI is next label below the top label being POPed.
- Max-payload size for the LSP gets reduced by 8 bytes after the insertion of the (ELI+EL).

4. Problems and Motivation

[EL-RFC] defines EL signaling/usage suitable for single-segment LSPs. However, as MPLS proliferates in the network access leading to the setup of e2e LSPs using LSP stitching and hierarchies, there is a need to define the EL behavior for LSP stitching and LSP hierarchies.

[EL-RFC] does not explicitly specify the EL-signaling-interaction between stitched LSPs segments. Similarly, peculiarities in the data-plane related to LSP stitching need further specification. Likewise, the signaling and data-plane peculiarities for using EL over LSP hierarchies could be further specified.

It is desirable to get the benefits of EL even for stitched LSPs.

Certain aspects peculiar to stitched LSPs need additional handling to increase the applicability of [EL-RFC]. [EL-RFC] needs to be extended

to define the behavior for LSP stitching and LSP hierarchies (tunneling) when using EL.

The sub-sections below list the specific additional requirements for making entropy label more deployable when used with LSP stitching, and LSP hierarchy.

4.1 EL applicability for seamless MPLS

The seamless MPLS architecture relies on the use of LSP stitching and hierarchy to signal an e2e LSP between access-nodes, such that the e2e LSP is going over aggregation/transport/cores nodes.

The signaling of such e2e LSPs is enabled by using the stitching/hierarchy mechanisms that exist, using [LU-BGP]/LDP/RSVP.

The rules of section 5 provide a general-purpose way for the use of ELs across e2e LSPs by defining:

- the rules of ELC propagation at stitching points;
- the data-plane guidelines for the stitching point LSR; and
- the data-plane guidelines for LSP hierarchies for inserting (ELI+EL) at ingress LER of an LSP in an LSP hierarchy.

4.2 EL for LSP stitching

A stitched e2e LSP might be stitched from greater than 2 segment LSPs (along the length of the e2e LSP), with 2 LSPs forming the stitch at each stitching point.

An LSP segment is considered to be intrinsically EL capable when it supports the attributes summarized in section 3.

Some of the segment LSPs in the e2e LSP may intrinsically support EL and some may not. So, the e2e LSP may not intrinsically support EL from end to end. However, to obtain the benefits of EL for stitched LSPs, it is important that an EL should be present on the data packets traversing as many segments of the e2e LSP as is possible within data plane abilities of the routers on the way.

In using EL with LSP stitching, the aims are BOTH of the following:

- a. Get EL benefits wherever possible: on all segments where possible. Just because a given segment does not support EL is not a reason to deny EL benefits to other segments of the e2e LSP.

- b. Not run into data-plane problems where an intermediate-segment whose ingress LER can not look deeper to remove EL when the subsequent segment does not support EL.

- Independent setup of LSP segments:

LSP stitching typically relies on LSP segments that have been independently setup. In an e2e LSP (made of stitched segments), it is unlikely that all of the stitching points (i.e., segment LSP end points) as well as the ultimate ingress and ultimate egress support EL as defined in section 3.

However, there would be individual LSP segments that would completely satisfy the requirements of section 2 (i.e. are intrinsically EL capable). This document describes how EL may be used for (portions of) the e2e LSP while still working within the framework for [EL-RFC].

S---A---B---C---D

In the above topology, for an e2e LSP from S to D, the segments AB and CD could be intrinsically EL capable while the segments SA, & BC may not be. For data traffic going over the LSP from S to D, using EL on the segments AB and CD would be beneficial for load-balancing over LAGs/ECMP.

- Dealing with different protocols being used to setup the segments of the e2e LSP.

4.2.1 Spectrum of EL usage behaviors required to be supported for stitched LSPs

To cater for an incremental deployment of intrinsically-ELC routers in a network, the multiple different modes for EL use with LSP stitching need to be to be supported.

The spectrum of supported behaviors are listed below by referencing the following diagram.

S1 S2 S3 S4

A-----B-----C-----D-----E

LSP segments S1, S2, S3, S4 are between LERs A/B/C/D/E. There may or may not be other routers between the per-segment ingress<->egress LERs.

Transport LSP signaling protocol: could be any: LDP/RSVP/([LU-BGP] tunneled over RSVP/LDP).

4.2.1.1 Entropy label for per-segment LSP

Each of the segments will have their independent EL capability based on BOTH the:

- Per-segment ingress having the ability to insert the EL.
- Per-segment egress (or PH router) having the ability to strip the EL.

This is very similar to [EL-RFC] with the additional data-plane rule of section 5.2.2.1 "A. Rationalizing EL for the outgoing LSP segment:".

Reasoning for why per-segment EL may be attractive for certain use scenarios:

Opportunity to get benefits on those segments where EL benefits are available. Even though the e2e LSP may not support ELC, this allows the EL benefits on those segments that are EL-capable.

4.2.1.2 Entropy label for notional-segment-LSP(s)

In the case of stitched LSPs, it is useful to:

- Insert EL at first per-segment ingress LER (per-segment ingress LER closest to the e2e ingress LER) that has the ability to insert EL.
- Carry the EL on the data packets as far along the stitched LSP as the last per-segment egress LER that ability to strip the EL on a series of contiguous EL-supporting segments.

The above is enabled by the rules of section "5.2.1.1 Signaling ELC at stitching points (Translation rules)".

The benefit of using EL with notional-segment LSPs:

An operator might be able to use EL for the MPLS traffic on its path to a stitching point even though the stitching-point router (or its PHR) itself may not have the data-plane capabilities required as in [EL-RFC].

Additionally, even if the stitching-point (or its PHR) do have the

data-plane capabilities of [EL-RFC], it is just more efficient to forward the data packets without having to strip the EL and then reinsert the EL when the downstream segment is also intrinsically ELC.

4.2.1.3 Entropy label for e2e LSP

This correspond to having the notional-LSP and the e2e LSP being the same.

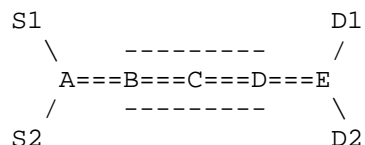
This is covered by the rules of section 5.2.1.1 "Signaling ELC at stitching points (Translation rules):" with the additional requirement that the data-plane be exactly the same as [EL-RFC], i.e.

the (ELI+EL) insertion is done by a label PUSHing router,
the (ELI+EL) POP is done by the PHR/egress for the e2e LSP.

4.3 EL for LSP hierarchy

For the purpose of highlighting the problem to be addressed and the resultant requirements to be met, the following diagram is presented as an example.

Let there be an LSP hierarchy with the ingress for the different levels of LSP hierarchy being at different routers, such that each LSP in the hierarchy is intrinsically EL capable. The individual LSPs in the hierarchy could be a single-segment LSP or a stitched e2e LSP.



In the above topology, let there be the following LSPs:

L1: B->D
L2: A->E, tunneled through LSP L1
L3: S1->D1, tunneled through LSP L2
L4: S2->D2, tunneled through LSP L2

All of the LSPs above are assumed to be intrinsically EL capable.

4.3.1 Possibility of unnecessary reduction of max-payload of the LSP:

Even though the aim of using EL is to get better load-balancing support, in some cases the insertion of (ELI+EL) may unnecessarily

reduce the effective payload of an LSP.

In above diagram, as per [EL-RFC] for a data packet on LSP L3, the insertion of (ELI+EL) for each of the 3 LSPs: L1, L2 and L3 is not explicitly prohibited. As a result it is possible that the packet on LSP L3, might end up with 3 (ELI+EL)s (one for each LSP level in the hierarchy) thus reducing the effective payload of the LSP L3. Likewise for L4. The presence of the (ELI+EL) for the outer LSPs L1 and L2 is not strictly useful for load-balancing the data traffic on the LSPs L3 and L4.

The solution for this issue is presented in section 5.3.2: it relies on inserting the (ELI+EL) in the context of only 1 LSP in a hierarchy.

This issues results in the following requirement for EL usage in the presence of LSP hierarchies:

- Desirability of having a single (ELI+EL) on data packets over an LSP hierarchy: The LSP for which the (ELI+EL) is inserted, is preferably the innermost intrinsically EL-capable LSP, as the notion of a user-flow is more specifically indicated by fields deeper inside the packet headers. Having an EL be present deeper in the packet provides load-balancing benefits of EL for the traversal of the packet across a longer stretch of the network.

If there is to be only 1 (ELI+EL) in the label stack, it imposes an additional data plane requirement on the ingress LER as described in section 5.3.2.

4.3.2 Possibility of EL being non-usable for load-balancing: Even though the aim of using EL is to get better load-balancing, in some cases the insertion of (ELI+EL) may actually offer no load-balancing benefits at all. Whether the presence of an EL offers load-balancing benefits on a given transit router, depends on:

- whether the transit router has a LAG or an ECMP as an outgoing interface for the LSP, AND
- whether the forwarding ASICs of the transit routers have the ability to include the EL (appearing at a specific position in the label stack) in the hash computation, either by:
 - + looking up the ELI and then picking the EL, or
 - + computing the hash on the maximum number of labels that it can pick from the label-stack for hash-computation which happens to also include the EL.

When the EL on a packet is outside the portion-of-the-label-stack that the ASIC of a transit router can use for hash computation, the forwarding hardware may include only the top few labels or the bottom

few labels in the hash computation. This may prevent the inclusion of EL for hash-computation.

In the above diagram, for a data packet going over LSP L3 let the issue of section 4.3.1 have been resolved by the router S1 inserting the (ELI+EL) underneath the label for LSP L3 and none of the other routers inserting the (ELI+EL). When this data packet arrives at router C, its label stack looks thus:

| | | | | | | | | |
|------------|--|------------|----|------------|--|-----|--|--------------|
| Label-LSP1 | | Label-LSP2 | | Label-LSP3 | | ELI | | EL |
| Top-label | | | -> | | | | | Bottom label |

Let's say that the router C is able to include only the top 4 labels in a label stack for the hash-computation due to the ability of its forwarding ASICs.

So, the router C is not able to get the benefit of the presence of the EL in the data packet. If the only ECMP/LAG in this network is the link between C&D, then the presence of the EL serves no purpose for the above network example and it ends up reducing the payload capacity of the LSPs L3 and L4 by 8 bytes.

This example could be further generalized in the case of seamless MPLS, where there may be deeper LSP hierarchies.

A transit router that has the ability to hash on an EL (based on its depth in the label stack) does not have multiple paths; while another router that has multiple paths and the ability hash on the EL (appearing at a specific depth in the label stack) is unable to do so as the EL appears outside the depth of the label stack that may be included in the hash. In neither of the foregoing cases is the presence of an EL helpful.

This translates into a requirement for EL: Flexibility in choice of LSP tunnel for which EL is inserted:

There is a need to have a way by which to include an EL underneath a specific label in a label-hierarchy based on it serving the most useful purpose (i.e. taking into consideration location of multiple-forwarding-paths and stack-depth-concerns).

[EL-RFC] has no way of influencing the insertion of (ELI+EL) at a certain LSP level in the stack. Thus, there is a need for a mechanism by which one of the many intrinsically-EL-capable LSPs in an LSP hierarchy could be picked for inserting the (ELI+EL).

5. EL for LSP stitching/hierarchy

5.1 Additional EL abstractions: specific to LSP stitching/hierarchy

Given the previous sections, following additional abstractions need to be defined to make EL more useful for LSP stitching and hierarchy.

5.2 New abstractions: EL applicability for LSP stitching

5.2.1 Signaling

New abstractions need to be defined to handle the differences in the use of ELs for stitched-LSPs as compared to their use for single-segment LSPs.

The differences are:

- Notion of ingress for EL insertion:
(ELI+EL) insertion might not necessarily be done by a label-PUSHing router. A stitching point where the label is being swapped might do the (ELI+EL) insertion, and serves as a "notional ingress".
- Notion of egress for EL:
"Notional-egress" might not be the segment egress for the segment of the notional-ingress.
Even though certain stitching-points (segment LERS) might not support POPing (ELI+EL), it may be acceptable to let the (ELI+EL) continue to be on the packet since the egress of a subsequent segment has the capability to POP the (ELI+EL) (which may not necessarily be along with POPing the transport label). A "notional-ingress and notional-egress" pair might actually be the segment-ingress and segment-egress for different LSP segments that are part of the same e2e LSP.

The portion of the stitched e2e LSP, between a notional-ingress and a notional-egress is referred to as the "notional-LSP-segment" in this document.

As a packet traverses an e2e LSP, it may have an (ELI+EL) imposed on it and then removed at different routers.

It is desirable for there to not be more than one instance of an (ELI+EL) to appear on a packet at any given time. However, the insertion followed by removal of an (ELI+EL) may happen more than once as the packet traverses the e2e LSP. Each router doing the (ELI+EL) insertion is the notional-ingress and each router doing the (ELI+EL) removal is the notional-egress (or notion EL-stripping-PH-router).

Thus, there may be more than 1 "notional ingress" for EL insertion, and there may be more than 1 "notional egress" for EL removal.

For each notional "ingress ingress", there will be a "notional egress" with the "notional ingress"es and "notional egress"es alternating along the path of the e2e LSP when there are more than 1 notional ingress and egress for an e2e LSP.

In the simplest case, this boils down to the case of there being just one notional ingress and one notional egress; and the notional ingress coincides with the e2e ingress, and the notional-egress coincides with the e2e egress. That is the case that [EL-RFC] addresses.

Separation of control/data-plane implies that certain routers

- Might be running software that supports signaling ELC and understanding an egress' ELC.
- However, might not have the capability to insert (ELI+EL).

Such routers should not be allowed to play a spoil-sport in preventing EL benefits for traffic traversing over them via stitched LSPs. In other words, such routers can not act as notional-ingress or notional-egress. However, the presence of such per-segment ingress/egress routers on the path of a notional segment-LSP should not prevent the notional segment-LSP from benefiting from the use of EL.

5.2.1.1 Signaling ELC at stitching points (Translation rules)

The rules for propagating ELC, at stitching points, from a downstream segment LSP to an upstream segment LSP are listed in this section.

There is benefit in propagating ELC across stitching points is to not have to re-compute the EL at different segment ingress for those segments that are EL capable, including when the LSP segments have been setup using different protocols.

Additionally, in certain cases it should be possible to get benefits of (ELI+EL) on LSP segments that are not "intrinsically EL capable", where the lack of "intrinsic EL capability" is due to:

- The per-segment ingress does not support EL insertion.
- The per-segment PHR/egress does not support EL POPing.

However, such a stitching point might support ELC signaling.

At a stitching point, when 2 LSP segments: L1 (incoming LSP) and L2 (the outgoing LSP) are being stitched, the following rules should be

followed by stitching point in signaling its ELC.

A. Segment-egress:

1. A segment-egress signals ELC for an LSP-segment L1 when:
 - a. The segment-egress is intrinsically ELC, or
 - b. When it is not intrinsically-ELC, however segment-egress for LSP-segment L2 (downstream of L1)- for which this stitching-point is segment-ingress - is signaling ELC.
[This handles the case: Support the signaling even though it may not support the data-plane behavior.]
2. A segment-egress MUST NOT signal ELC if BOTH of the following are true:
 - a. It is also segment-ingress for another LSP-segment whose segment-egress is not signaling ELC.
 - b. This router does not have the ability to remove an (ELI+EL) inserted by the segment-ingress for the LSP-segment for which this router is the segment-egress.

B. Segment-ingress:

The following is relevant only for RSVP as defined in [EL-RFC]. When this router acting as segment-egress for an LSP L1 (that is stitched to downstream LSP L2) is signaling ELC for L1, then this router must signal ELC in its Path messages using the mechanism defined in [EL-RFC].

This is relevant only in the context of bidirectional LSPs.

5.2.2 Data plane aspects

5.2.2.1 Stitching: Differing EL dispositions

At a stitching point, when 2 LSP segments: L1 (incoming LSP) and L2 (the outgoing LSP) are being stitched, the following rules should be followed by the stitching point in its data-plane behavior.

A. Rationalizing EL for the outgoing LSP segment:

When the LSP segments L1 and L2 differ in their ELC, the stitching point router needs to take the following data-plane actions depending on its role for the e2e LSP:

a. Notional egress behavior:

When L1 intrinsically supports ELC and L2 does not, then the stitching-point router must remove the (ELI+EL), if present under top label, from the received data packets before effectively SWAPing the top label. In other words,

in the presence of the ELI, the operations performed should be:

```
POP(IncomingLabel), POP(ELI+EL), PUSH(OutgoingLabel)
    or alternately:
POP, POP, SWAP(OutgoingLabel)
```

Translation rule "A 2" of section 5.2.1.1 would have ensured that the above is doable at the stitching point.

b. Notional ingress behavior:

When L1 does not intrinsically support ELC and L2 does, then the stitching point router must POP the incoming label, insert (ELI+EL) before PUSHing the label for the LSP segment L2.

The label operations performed would be:

```
POP(IncomingLabel), PUSH(EL), PUSH(ELI), PUSH(OutgoingLabel),
    or
SWAP(EL), PUSH(ELI), PUSH(OutgoingLabel)
```

c. Implicit notional ingress behavior:

When L1 is intrinsically ELC and so is L2, the arriving data traffic should already have (ELI+EL) on it.

However, it is possible that due to local configuration on the notional-ingress, (ELI+EL) is not being inserted. In that case, traffic arriving on L1 will not have (ELI+EL) on it.

In that case, this stitching-point is the "implicit notional ingress" and it should insert (ELI+EL) just as if it were a "notional ingress".

B. Preventing multiple (ELI+EL) pairs underneath a given forwarding label in the stack:

A segment-ingress that is intrinsically-EL-capable should have the ability to inspect received data packets to check whether an (ELI+EL) already exists on the data packet underneath the top label.

Not causing multiple ELs on a data packet:

When both the LSP segments L1 and L2 support ELC, the stitching point router SHOULD insert an (ELI+EL) only if the incoming packet does not contain an (ELI+EL) underneath the top label. In that case, the label operations are as below:

```
POP(IncomingLabel), PUSH(ELI+EL), PUSH(OutgoingLabel)
```


If the incoming packet already contains an (ELI+EL) underneath the top label, an additional (ELI+EL) MUST NOT be inserted on the packet underneath the top label that is being effectively SWAPed.

This prevents a segment ingress from inserting an (ELI+EL) when the notional ingress has already inserted an (ELI+EL).

C. Rationalizing EL insertion (at stitching-point) for LSP hierarchy:

A stitching point router that is intrinsically-EL-capable should have the ability to inspect received data packets to check whether an (ELI+EL) already exists, underneath any label in the label-stack.

If the router has such a ability, then this router MUST NOT insert an (ELI+EL) as in "A a" above.

This helps to prevent multiple (ELI+EL)s on the packet inserted (at a stitching point) in the context of different transport labels in a label hierarchy.

D. Notional ingress role change at a router:

This role can change due to local configuration on the router or due to segment egress starting/stopping to signal ELC possibly due to a configuration change at the segment egress or due to a configuration change at this router. When this router becomes a notional ingress, it reacts to the change as in "A b" above.

When this router stops being a notional ingress, this router stops inserting the (ELI+EL) underneath the top label that this router is

SWAPing (if this router is stitching point), or
PUSHING (if this router is e2e ingress).

E. Notional egress role change at a router:

This role can change due to local configuration on the router or due the egress of a downstream stitched LSP segment starting to signal ELC.

When this router becomes a notional egress, it reacts to the change as in "A a" above.

When this router stops being a notional egress, this router stops

performing the label operation described in "A a" above. Instead this router now starts to simply SWAP the top label.

5.3 New abstractions: EL applicability for LSP hierarchy

5.3.1 Management plane:

Moving the (ELI+EL) underneath a different LSP's transport label:

There are 2 ways to handle the issue of section 4.3.2:

- Configuration at the ingress LER: a configuration option should exist by which an operator can disable the insertion of (ELI+EL) on a per-LSP basis. The specific level in the LSP hierarchy for which to enable this configuration is based on operator knowledge based on:
 - * Knowledge of transit routers that need EL benefits : those routers that have a multi-path (LAG or LSP ECMP) as egress interface.
 - * The label hashing abilities of such routers: information about the specific number of labels in the label-stack that can be used in the hash computation; and any constraints about the position of the labels that can be used for computation when the label stack is larger than a certain ASIC-specific threshold.
- Configuration-based rewrite of the label stack at the ingress LER of an intrinsically-EL-capable LSP:

An operator will know the forwarding characteristics (with regards to the number of labels that can be included in the hash computation) of the transit routers across the path of the e2e LSP that is part of an LSP hierarchy.

By making such a configuration, the operator can ensure that the EL will appear in the label stack such that all transit routers shall be able to include the as part of the hash computation.

The configuration would cause the label stack of the outgoing packet to have its extant (ELI+EL) removed, and an (ELI+EL) inserted just underneath the label of the LSP for which this ingress LER is setup to insert (ELI+EL).

5.3.2 Data plane aspects

Preventing insertion of multiple (ELI+EL)s:

At an ingress LER, the router SHOULD not insert an (ELI+EL) for an LSP if the packet already contains an ELI.

This ensures that for a data packet on a hierarchy of LSPs, there will be only 1 instance of an (ELI+EL). This helps to prevent the issue of section 4.3.1.

This also ensures that when multiple LSPs in an LSP hierarchy are intrinsically-EL-capable, the (ELI+EL) will be inserted just underneath the transport label of the innermost LSP in the hierarchy. However, based on section 5.3.1 there is a way by which to modify the level in the LSP hierarchy for which an (ELI+EL) is inserted.

A more specific case of this is already covered in section "5.2.2.1 C. Rationalizing EL insertion (at stitching-point) for LSP hierarchy:".

6. Security considerations

Security considerations as listed in section 9 of [EL-RFC] apply.

7. Acknowledgments

Many thanks to Adrian Farrel for his inputs on the stitching scenarios, and suggesting editorial improvements.

Thanks to the EL team (Sudharsana Venkataraman, Nitin Singh, Ramji Vijayaraghavan, Jie Yan, Abhishek Tripathi) for discussions on some of these topics.

8. IANA considerations

None.

9 References

9.1 Normative References

- [EL-RFC] Kompella, K., Drake, J., Amante, S., Henderickx, W., L. Yong, "The Use of Entropy Labels in MPLS Forwarding", RFC-6790, November 2012.

- [GMPLS-HIER] Kompella, K., Y. Rekhter, "Label Switched Paths (LSP) Hierarchy with Generalized Multi-Protocol Label Switching (GMPLS)", RFC-4206, October 2005.
- [MPLS-ARCH] Rosen, E., Viswanathan, A., R. Callon, "Multiprotocol Label Switching Architecture", RFC-3031, January 2001.
- [S-MPLS] Leymann, N., Decraene, B., Filsfils, C., Konstantynowicz, M., Steinberg, D., "Seamless MPLS Architecture", draft-leymann-mpls-seamless-mpls, October 2012.
- [STITCHING] Ayyangar, A., Kompella, K., Vasseur, JP., A. Farrel, "Label Switched Path Stitching with Generalized Multiprotocol Label Switching Traffic Engineering (GMPLS TE)", RFC 5150, February 2008.

9.2 Informative References

- [ISSUE-DEEP] K. Kompella, "Deep Label Stacks", <http://tools.ietf.org/agenda/84/slides/slides-84-mpls-15.pdf>, August 2012
- [LU-BGP] Rekhter, Y., E. Rosen, "Carrying Label Information in BGP-4", RFC-3107, May 2001.

Authors' Addresses

Ravi Singh
Juniper Networks
1194 N. Mathilda Ave.
Sunnyvale, CA 94089
US

EMail: ravis@juniper.net

Yimin Shen
Juniper Networks
10 Technology Park Drive
Westford, MA 01886
US

EMail: yshen@juniper.net

John Drake
Juniper Networks
1194 N. Mathilda Ave.
Sunnyvale, CA 94089
US

EMail: jdrake@juniper.net

Internet Engineering Task Force
Internet-Draft
Intended status: Standards Track
Expires: August 12, 2013

Y. Shen
Juniper Networks
Y. Kamite
NTT Communications Corporation
February 8, 2013

RSVP Setup Protection
draft-shen-mpls-rsvp-setup-protection-02

Abstract

RFC 4090 specifies an RSVP facility-backup fast reroute mechanism for protecting established LSPs against link and node failures. This document extends the mechanism to provide so-called "setup protection" for LSPs during their initial Path message signaling time. In particular, it enables a router to reroute an LSP via an existing bypass LSP, when there is a failure of the immediate downstream link or node along the desired path. Therefore, it can be used to reduce LSP setup time in such a situation, or allow LSPs with strict paths to be established successfully when alternative paths are unavailable in the network or unable to be computed by ingress.

Status of this Memo

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

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

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

This Internet-Draft will expire on August 12, 2013.

Copyright Notice

Copyright (c) 2013 IETF Trust and the persons identified as the document authors. All rights reserved.

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

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

Table of Contents

| | |
|--|----|
| 1. Introduction | 3 |
| 2. Specification of Requirements | 4 |
| 3. Theory of Operation | 4 |
| 3.1. New RSVP Attribute Flag | 5 |
| 3.2. New RSVP Attributes TLVs | 5 |
| 3.2.1. Protected LSP Sender IPv4 Address TLV | 6 |
| 3.2.2. Protected LSP Sender IPv6 Address TLV | 6 |
| 3.3. PLR behavior | 7 |
| 3.4. MP behavior | 9 |
| 3.5. Local Revertive Mode | 10 |
| 4. IANA Considerations | 10 |
| 5. Security Considerations | 10 |
| 6. Acknowledgements | 10 |
| 7. References | 10 |
| 7.1. Normative References | 10 |
| 7.2. Informative References | 11 |
| Authors' Addresses | 11 |

1. Introduction

In RSVP facility-backup fast reroute (FRR) [RFC 4090], the router at a point of local repair (PLR) of an LSP can redirect traffic via a bypass LSP upon a failure of the immediate downstream link or node. Such protection is normally established after the LSP has been set up. This is because the PLR must know the label and address of the next-hop router (in the case of link protection) or those of the next-next-hop router (in the case of node protection), before it can select or signal a bypass LSP to protect the LSP. The information of the label and the address is carried in a Resv message.

Imagine a scenario where a new LSP is being signaled, but its Path message carries an EXPLICIT_ROUTE object (ERO) with a strict path that is statically configured or computed offline based on a topology that assumes no failure in the network. In such a case, if a link or node along the path happens to be in a failure condition, RSVP signaling will stop at the router upstream adjacent to the failure. This will be the case even if there is an existing bypass LSP protecting the link or node for some existing LSPs. In other words, this new LSP is not protected during this setup phase, i.e. the initial Path message signaling time.

In this situation, the network would normally rely on IGP to update traffic engineering (TE) information throughout the network, and the router upstream adjacent to the failure to send a PathErr message to trigger the ingress router to compute and signal a new path. However, this approach may not always be possible, desirable, or even relevant in the following scenarios:

1. Static strict path. As described above, if the ERO carries an explicit path with a sequence of strict hops that are statically configured or computed offline based on a topology assuming no network failure, the LSP will never be established.
2. LSPs with a strict requirement for setup time. IGP TE information flooding, PathErr message propagation, and path re-computation and re-signaling may introduce a significant delay to LSP establishment. This may impact on the setup time of services that have a strict requirement for it, such as on-demand transport services for real-time data.
3. Sibling P2MP sub-LSPs sharing a common link. In this case, the new LSP is a sub-LSP of a P2MP LSP, and its desired path is supposed to share the failed link with an existing sibling sub-LSP, i.e. another sub-LSP of the same P2MP LSP, which is being protected by a bypass LSP. If the new sub-LSP is rerouted via a different path, it will not be able to share the data flow over

the bypass LSP with that sibling sub-LSP, creating unnecessary traffic flow in the network.

This document extends the RSVP facility-backup fast reroute mechanism to provide so-called "setup protection". During the initial Path message signaling of an LSP, if there is a link or node failure along the desired path, and if there is a bypass LSP protecting the link or node, the LSP will be signaled through the bypass LSP. The LSP will be established as if it was originally set up along the desired path (aka. primary path) and then failed over to the bypass LSP after the failure. After the failure is resolved, the LSP MAY be reverted to the primary path. The mechanism is applicable to both P2P and P2MP LSPs.

2. Specification of Requirements

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

3. Theory of Operation

When an LSP is being signaled by RSVP, a Path message is sent hop by hop from the ingress router to the egress router, following the path defined by an ERO. The setup protection mechanism in this document enables an ingress or transit router to reroute the LSP via a bypass LSP, if the router detects a failure of the immediate downstream link or node represented by the next hop in the ERO, called "next ERO hop". In this case, the current router is referred to as a PLR.

The mechanism is relevant when the Path message carries the "local protection desired" flag in the SESSION_ATTRIBUTE object [RFC 4090] and a new "setup protection desired" flag defined in this document (Section 3.1).

On a PLR, the mechanism is only applicable when the next ERO hop is a strict hop, and in the case of node protection, the next-next ERO hop is also a strict hop. A strict next ERO hop allows the PLR to unambiguously decide the intended downstream link or node along the desired path, and hence reliably detect its status. In link protection, the strict next ERO hop also indicates the merge point (MP), i.e. the destination of the bypass LSP to be used to reroute the LSP. In node protection, the strict next-next ERO hop indicates the MP.

When performing setup protection, the PLR signals a backup LSP by

tunneling Path message through the bypass LSP. Like the Path message of a backup LSP in the normal facility-backup FRR ([RFC 4090]), this Path message carries an address of the PLR as the sender address in SENDER_TEMPLATE object. In addition, the Path message also carries the information of the protected LSP (Section 3.2). When the MP receives the Path message, it terminates the backup LSP, and re-creates the protected LSP. If the MP is the egress router of the protected LSP, it terminates the protected LSP as well. If the MP is a transit router of the protected LSP, it signals the LSP further downstream.

Eventually, the LSP will be established end to end, with the backup LSP tunneled through the bypass LSP from the PLR to the MP. The RSVP state on the PLR and the MP and the RSVP messages generated by these routers are no different than those in a post-failure situation of a normal facility-backup FRR.

Later, when the failure is resolved, the PLR MAY revert the LSP to the primary path, in the same manner as the local revertive mode specified in [RFC 4090].

The setup protection MAY be enabled and disabled on a router based on configuration. For an LSP to be setup-protected, the mode MUST be enabled on both PLR and MP. If it is enabled on the PLR but disabled on the MP, the MP SHOULD reject the Path message of the backup LSP and send a PathErr message, as described Section 3.4.

3.1. New RSVP Attribute Flag

In order for an LSP to explicit request for setup protection, this document defines a new "setup protection desired" flag in the Attribute Flags TLV of the LSP_ATTRIBUTES object [RFC5420]. It is carried in the Path message of the LSP, i.e. the protected LSP.

3.2. New RSVP Attributes TLVs

This document defines two new RSVP Attributes TLVs [RFC 5420]. They are used by a PLR to convey to an MP the original sender address in the SENDER_TEMPLATE object of a protected LSP. Both TLVs are carried in the LSP_REQUIRED_ATTRIBUTES object in the Path message of a backup LSP.

- o Protected LSP Sender IPv4 Address TLV
- o Protected LSP Sender IPv6 Address TLV

3.2.1. Protected LSP Sender IPv4 Address TLV

The Protected LSP Sender IPv4 Address TLV is defined with type TBD. It is allowed on LSP_REQUIRED_ATTRIBUTES object, and not allowed on LSP_ATTRIBUTES object. The encoding is as below.

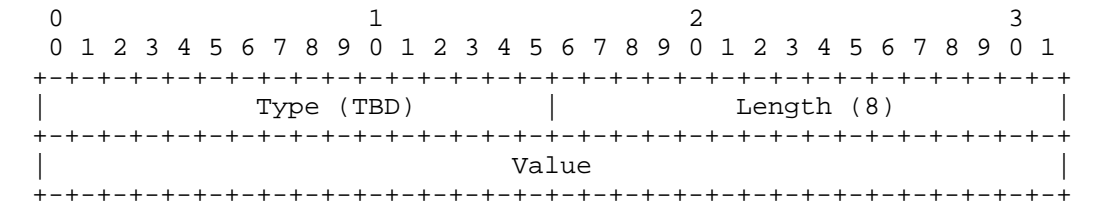


Figure 1

Type

TBD

Length

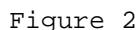
8

Value

Original sender address in the IPv4 SENDER_TEMPLATE object of the protected LSP.

3.2.2. Protected LSP Sender IPv6 Address TLV

The Protected LSP Sender IPv6 Address TLV is defined with type TBD. It is allowed on LSP_REQUIRED_ATTRIBUTES object, and not allowed on LSP_ATTRIBUTES object. The encoding is as below.



If a bypass LSP is not found by the above criteria, the router MUST originate a PathErr with code = 24 (routing problem) and sub-code = 2 (bad strict node).

If a bypass LSP is found, the router MUST act as a PLR for setup protection, and reroute the protected LSP via the bypass LSP. If multiple satisfactory bypass LSPs exist, the PLR MAY select one based on bandwidth constraints or local policies. Specifically, if the protected LSP is a sub-LSP of a P2MP LSP, a bypass LSP that is protecting an existing sibling sub-LSP MUST be preferred, in order to minimize traffic duplication in the network.

The PLR SHOULD NOT send the Path message of the protected LSP any further. Instead, it MUST create a backup LSP, and send a Path message of the backup LSP to the MP via the bypass LSP. The Path message is constructed by using the sender template specific method [RFC 4090]. In particular, it has the sender address in the SENDER_TEMPLATE object set to an address of the PLR. It MUST also carry an LSP_REQUIRED_ATTRIBUTES object with a Protected LSP Sender IPv4 Address TLV or Protected LSP Sender IPv6 Address TLV.

Upon receiving a Resv message of the backup LSP from the MP, the PLR SHOULD bring up both of the backup LSP and the protected LSP. If the PLR is the ingress router of the protected LSP, the LSP has been set up successfully. If the PLR is a transit router, it MUST send a Resv message upstream for the protected LSP, with the "local protection available", "local protection in use", and optionally "node protection" and "bandwidth protection" flags set to 1, in the RRO hop corresponding to the PLR [RFC 4090]. The PLR SHOULD originate a PathErr message with code = 25 (notify error) and sub-code = 3 (tunnel locally repaired).

The PLR SHOULD also install a forwarding entry for the protected LSP. In the typical case, the forwarding entry should result in two outgoing labels for packets. The inner label is the backup LSP's label, and the outer label is the bypass LSP's label. However, the forwarding entry may result in one or no label, if either or both of the backup LSP and the bypass LSP have the Implicit NULL label.

If the PLR receives a PathErr message when signaling the backup LSP, the PLR MUST NOT bring up the backup LSP or the protected LSP. If the PLR is a transit router of the protected LSP, it MUST send a PathErr message upstream for the protected LSP. Likewise, if the PLR receives a PathErr message of the backup LSP after the backup LSP and the primary LSP have previously been brought up, and the PLR is a transit router of the protected LSP, it MUST also send a PathErr message upstream for the protected LSP.

When the PLR receives a ResvTear message of the backup LSP, the PLR MUST bring down both the backup LSP and the protected LSP. If the PLR is a transit router of the protected LSP, it MUST send a ResvTear message upstream for the protected LSP.

In any cases where the PLR needs to bring down the protected LSP due to a received PathTear message, an RSVP state time-out, a configuration change, an administrative command, etc, the PLR MUST also bring down the backup LSP by sending a PathTear message through the bypass LSP.

3.4. MP behavior

When an MP receives the Path message of a backup LSP, it MUST realize the setup protection condition based on the presence of Protected LSP Sender IPv4 Address TLV or Protected LSP Sender IPv6 Address TLV in LSP_REQUIRED_ATTRIBUTES object.

If setup protection mode is disabled on the MP, it MUST reject the Path message, by sending a PathErr with code = 2 (policy control failure) to the PLR.

Otherwise, the MP MUST terminate the backup LSP and re-create the protected LSP. If the MP is the egress router of the protected LSP, it MUST also terminate the protected LSP. If the MP is a transit router of the LSP, it MUST send a Path message downstream for the protected LSP. The Path message has the sender address in SENDER_TEMPLATE object set to the original address of the ingress router, based on the above received Protected LSP Sender IPv4 Address TLV or Protected LSP Sender IPv6 Address TLV. The Path message MUST NOT carry any Protected LSP Sender IPv4 Address TLV or Protected LSP Sender IPv6 Address TLV in LSP_REQUIRED_ATTRIBUTES object.

The MP MUST allocate a label for the backup LSP, and distribute it to the PLR via Resv message of the backup LSP. If the protected LSP is a sub-LSP of a P2MP LSP and there is an existing sibling sub-LSP whose backup LSP is tunneled through the same bypass LSP, the MP MUST allocate the same label as the sibling sub-LSP, in order to avoid traffic duplication at the PLR.

When the MP receives a PathTear message for the backup LSP, it MUST bring down both the backup LSP and the protected LSP. If the MP is a transit router of the protected LSP, it MUST send a PathTear message downstream for the protected LSP.

In any cases where the MP receives or originates a PathErr or ResvTear message for the protected LSP, the MP MUST translate the message to a same type of message for the backup LSP and send it to

the PLR.

3.5. Local Revertive Mode

When the failed link or node is restored, the PLR MAY revert the protected LSP to its desired primary path, by following the procedure of local revertive mode described in [RFC 4090].

4. IANA Considerations

This document defines a new flag for the Attribute Flags TLV, which is carried in the LSP_ATTRIBUTES Object of Path message. This flag is used to communicate whether setup protection is desired for an LSP. The value of the new flag needs to be assigned by IANA.

Setup Protection Desired: TBD

This document defines two new RSVP Attributes TLVs, which are carried in the LSP_REQUIRED_ATTRIBUTES object of Path message. The values of the new types need to be assigned by IANA.

Protected LSP Sender IPv4 Address TLV

Protected LSP Sender IPv6 Address TLV

5. Security Considerations

The security considerations discussed in RFC 3209, RFC 4090 and RFC 4875 apply to this document.

6. Acknowledgements

Thanks to Rahul Aggarwal, Disha Chopra, and Nischal Sheth for their contribution.

7. References

7.1. Normative References

[RFC2205] Braden, B., Zhang, L., Berson, S., Herzog, S., and S. Jamin, "Resource ReSerVation Protocol (RSVP) -- Version 1 Functional Specification", RFC 2205, September 1997.

[RFC3209] Awduche, D., Berger, L., Gan, D., Li, T., Srinivasan, V.,

and G. Swallow, "RSVP-TE: Extensions to RSVP for LSP Tunnels", RFC 3209, December 2001.

- [RFC4090] Pan, P., Swallow, G., and A. Atlas, "Fast Reroute Extensions to RSVP-TE for LSP Tunnels", RFC 4090, May 2005.
- [RFC5420] Farrel, A., Papadimitriou, D., Vasseur, JP., and A. Ayyangarps, "Encoding of Attributes for MPLS LSP Establishment Using Resource Reservation Protocol Traffic Engineering (RSVP-TE)", RFC 5420, February 2009.
- [RFC4875] Aggarwal, R., Papadimitriou, D., and S. Yasukawa, "Extensions to Resource Reservation Protocol - Traffic Engineering (RSVP-TE) for Point-to-Multipoint TE Label Switched Paths (LSPs)", RFC 4875, May 2007.
- [RFC3471] Berger, L., "Generalized Multi-Protocol Label Switching (GMPLS) Signaling Functional Description", RFC 3471, January 2003.
- [RFC3472] Ashwood-Smith, P. and L. Berger, "Generalized Multi-Protocol Label Switching (GMPLS) Signaling Constraint-based Routed Label Distribution Protocol (CR-LDP) Extensions", RFC 3472, January 2003.
- [RFC3031] Rosen, E., Viswanathan, A., and R. Callon, "Multiprotocol Label Switching Architecture", RFC 3031, January 2001.

7.2. Informative References

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

Authors' Addresses

Yimin Shen
Juniper Networks
10 Technology Park Drive
Westford, MA 01886
USA

Phone: +1 9785890722
Email: yshen@juniper.net

Yuji Kamite
NTT Communications Corporation
Granpark Tower 3-4-1 Shibaura, Minato-ku
Tokyo 108-8118
Japan

Email: y.kamite@ntt.com

CCAMP Working Group
Internet-Draft
Intended Status: Standards Track
Expires: April 15, 2013

Mike Taillon
Tarek Saad
Rakesh Gandhi
Zafar Ali
Cisco Systems, Inc
October 12, 2012

Extensions to Resource Reservation Protocol For Fast Reroute of
Bidirectional Co-routed Traffic Engineering LSPs
draft-tsaad-ccamp-rsvpte-bidir-lsp-fastreroute-00

Abstract

This document defines RSVP-TE signaling extensions to support Fast Reroute (FRR) of bidirectional co-routed Traffic Engineering (TE) LSPs. These extensions enable the re-direction of bi-directional traffic and signaling onto bypass tunnels that ensure co-routedness of data and signaling paths in the forward and reverse directions after FRR. In addition, the RSVP-TE signaling extensions allow the coordination of bypass tunnel assignment protecting a common facility in both forward and reverse directions prior to or post failure occurrence.

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 | 3 |
| 2. Terminology | 3 |
| 3. Link Failure With Node-protection Bypass Tunnels | 5 |
| 3.1. Behavior Before Local Repair | 5 |
| 3.1.1. Downstream Merge Point Label Discovery | 6 |
| 3.1.2. Upstream Merge Point Label Discovery | 6 |
| 3.2. Behavior Post Link Failure After FRR | 6 |
| 3.3. Behavior Post Link Failure To Re-coroute | 6 |
| 4. Bypass Tunnel Assignment | 8 |
| 4.1. DOWNSTREAM_BYPASS_ASSIGNMENT Object | 8 |
| 4.2. Bypass Tunnel Assignment Signaling Procedure | 10 |
| 5. Compatibility | 10 |
| 6. Security Considerations | 10 |
| 7. IANA Considerations | 11 |
| 8. Acknowledgements | 11 |
| 9. References | 11 |
| 9.1. Normative References | 11 |
| Authors' Addresses | 12 |

1. Introduction

Co-routed bidirectional tunnels are signaled using GMPLS signaling procedures specified in [RFC3473] and [RFC3471]. Existing procedures defined in [RFC4090] describe the behavior of the Point of Local Repair (PLR) to reroute traffic and signaling onto the bypass tunnel in the event of a failure for unidirectional LSPs. These procedures are applicable to unidirectional protected LSPs, and don't address issues that arise employing FRR for bidirectional co-routed Label Switched Paths (LSPs).

When using current FRR procedures with bidirectional co-routed LSPs, it is possible in some cases (e.g. when using node-protecting bypass tunnels post a link failure event and when RSVP signaling is sent in-fiber and in-band with data), the RSVP signaling refreshes may stop reaching some nodes along the primary bidirectional LSP path after the PLRs complete rerouting traffic and signaling onto the bypass tunnels. This is caused by the asymmetry of paths that may be taken by the bidirectional LSP's signaling in the forward and reverse directions after FRR reroute. In such cases, the RSVP soft-state timeout eventually causes the protected bidirectional LSP to be destroyed, and consequently impacts protected traffic flow after FRR. This problem exists when using either unidirectional or bidirectional bypass tunnels to protect the primary co-routed bidirectional LSP.

When co-routed bidirectional bypass tunnels are used to locally protect bidirectional LSPs, the upstream and downstream PLRs may independently assign different bidirectional bypass tunnels in the forward and reverse direction. Currently, there is no means to coordinate the bypass tunnel selection between the downstream and upstream PLRs. In case of mismatch and after FRR, data traffic and signaling may flow over asymmetric paths in the forward and reverse directions which may be undesirable for certain applications.

This document proposes solutions to the above problems by providing corrective actions in the control plane to complement FRR procedures of [RFC4090] in order to maintain the RSVP soft-state for bidirectional protected LSPs and achieve symmetry in the paths followed by data and signaling in the forward and reverse directions post FRR. The document also extends RSVP signaling so it is possible that the bypass tunnel selected by the upstream PLR matches the one selected by the downstream PLR.

2. Terminology

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this

document are to be interpreted as described in RFC 2119 [RFC2119].

The reader is assumed to be familiar with the terminology in [RSVP] and [RSVP-TE].

LSR: Label-Switch Router.

LSP: An MPLS Label-Switched Path. In this document, an LSP will always be explicitly routed.

Local Repair: Techniques used to repair LSP tunnels quickly when a node or link along the LSP's path fails.

PLR: Point of Local Repair. The head-end LSR of a backup tunnel or a detour LSP.

Facility Backup: A local repair method in which a bypass tunnel is used to protect one or more protected LSPs that traverse the PLR, the resource being protected, and the Merge Point in that order.

Protected LSP: An LSP is said to be protected at a given hop if it has one or multiple associated backup tunnels originating at that hop.

Bypass Tunnel: An LSP that is used to protect a set of LSPs passing over a common facility.

Backup Tunnel: The LSP that is used to backup up one of the many LSPs in many-to-one backup.

NHOP Bypass Tunnel: Next-Hop Bypass Tunnel. A backup tunnel that bypasses a single link of the protected LSP.

NNHOP Bypass Tunnel: Next-Next-Hop Bypass Tunnel. A backup tunnel that bypasses a single node of the protected LSP.

Backup Path: The LSP that is responsible for backing up one protected LSP. A backup path refers to either a detour LSP or a backup tunnel.

MP: Merge Point. The LSR where one or more backup tunnels rejoin the path of the protected LSP downstream of the potential failure. The same LSR may be both an MP and a PLR simultaneously.

CSPF: Constraint-based Shortest Path First.

Downstream PLR: A PLR that locally detects a fault and reroutes traffic in the same direction of the protected bidirectional LSP RSVP Path signaling.

Upstream PLR: A PLR that locally detects a fault and reroutes traffic in the opposite direction of the protected bidirectional LSP RSVP Path signaling.

Point of Remote Repair (PRR): an upstream PLR that triggers reroute of traffic and signaling based on procedures described in this document.

3. Link Failure With Node-protection Bypass Tunnels

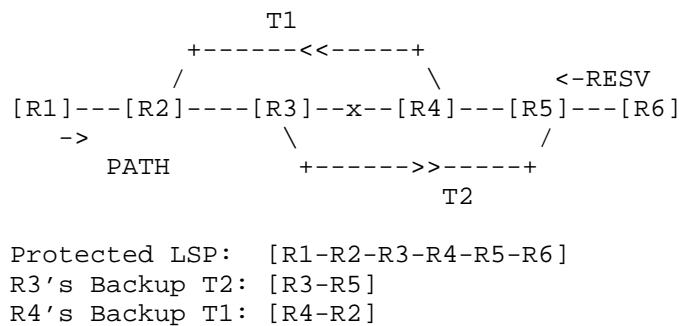


Figure 1: Flow of RSVP signaling post FRR after failure

Consider the Traffic Engineered (TE) network shown in Figure 1. Assume every link in the network is protected with a node- protection bypass tunnel. For the protected bidirectional co-routed LSP whose active/head is on router R1 and passive/tail is on router R6, each traversed router (a potential PLR) independently assigns a node- protection bypass tunnel. Consider a link R3-R4 on the LSP path fails.

The proposed solution introduces two phases to invoking FRR procedures by the PLR post the link failure. The first phase comprises of FRR procedures to fast reroute data traffic onto bypass tunnels in the forward and reverse direction. The second phase re- coroutes the data and signaling in cases where they go over asymmetric paths in the forward and reverse directions after the first phase.

3.1. Behavior Before Local Repair

To correctly reroute data traffic over a node-protection tunnel, the downstream and upstream PLRs have to know, in advance, the downstream and upstream Merge Point (MP) labels so that data in the forward and

reverse directions can be tunneled through the bypass tunnel post FRR respectively.

3.1.1. Downstream Merge Point Label Discovery

For unidirectional primary LSPs, [RFC4090] defines procedures for the downstream PLR to obtain the downstream MP label from recorded labels of the RSVP Resv message received at the downstream PLR.

3.1.2. Upstream Merge Point Label Discovery

To obtain the upstream MP label, existing methods to record upstream MP label in the RRO of the RSVP Path message are used. The upstream PLR can obtain the upstream MP label from the recorded label in the RRO of the received RSVP Path message.

3.2. Behavior Post Link Failure After FRR

The downstream PLR R3 and upstream PLR R4 independently trigger fast reroute procedures to redirect traffic onto respective bypass tunnels T2 and T1 in the forward and reverse direction. The downstream PLR R3 also reroutes RSVP Path state onto the bypass tunnel T2 using procedures described in [RFC4090]. Note, at this point, router R4 stops receiving RSVP Path refreshes for the protected bidirectional LSP while primary protected traffic continues to flow over bypass tunnels.

3.3. Behavior Post Link Failure To Re-coroute

The downstream Merge Point (MP) R5 that receives rerouted protected LSP RSVP Path message through the bypass tunnel, in addition to the regular MP processing defined in RF4090, gets promoted to a Point of Remote Repair (PRR role) and performs the following actions to re-coroute signaling and data traffic over the same path in both directions:

For unidirectional bypass tunnels:

- Checks for presence of a bypass tunnel in the reverse direction that terminates on the Downstream PLR R3. Note: the Downstream PLR R3's address is extracted from the "IPV4 tunnel sender address" in the SENDER_TEMPLATE object.
- If present, checks whether the primary LSP traffic and signaling is already rerouted over the found bypass tunnel. If not, PRR R5 activates FRR reroute procedures to direct traffic and signaling (RSVP Resv) over the found bypass tunnel T3 in reverse

direction.

- If not present, PRR R5 attempts to auto-provision a bypass tunnel that terminates on the downstream PLR R3. For unidirectional bypass tunnels, if co-routedness in forward and reverse direction is desired, the reverse path bypass tunnel can be inferred from the forward bypass tunnel path (e.g. by reflecting the RRO recorded in the forward direction as ERO for the reverse direction).
- If PRR R5 is unable to successfully provision a bypass tunnel that terminates on the downstream PLR, it may send an immediate RSVP Notify message back to the head-end. The head-end may tear and re-setup the LSP immediately.

For bidirectional bypass tunnels:

- The PRR follows similar procedures described in the solution to second problem in order to identify the bypass tunnel, and reroute traffic and signaling in the reverse path.

If MP R5 receives multiple RSVP Path messages through multiple bypass tunnels (e.g. as a result of multiple failures), the PRR should identify/provision a bypass tunnel that terminates on the farthest downstream PLR along the protected LSP path (closest to the bidirectional tunnel headend) and activate the reroute procedures mentioned above.

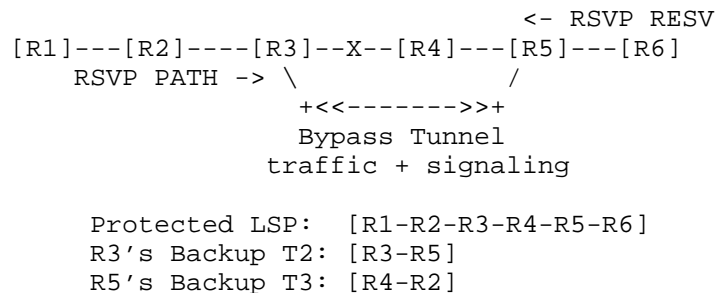


Figure 2: Flow of RSVP signaling post FRR after re-coroute

Figure 2 describes the path taken by traffic and signaling after completing re-coroute of data and signaling in the forward and reverse paths described earlier.

4. Bypass Tunnel Assignment Coordination

This document defines one additional RSVP object, DOWNSTREAM_BYPASS_ASSIGNMENT, to extend RSVP-TE for fast-reroute signaling. This object is backward compatible with LSRs that do not recognize it (see section 3.10 in [RSVP]).

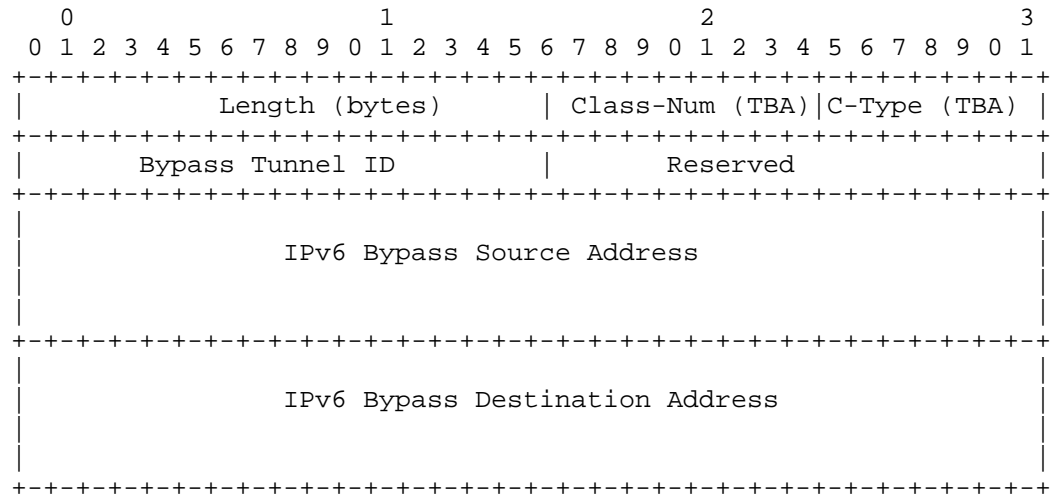
4.1. DOWNSTREAM_BYPASS_ASSIGNMENT Object

The DOWNSTREAM_BYPASS_ASSIGNMENT object is used to coordinate the backup used for the protected LSP by the downstream and upstream PLRs in the forward and reverse direction respectively prior or post the failure occurrence. This object MUST only be inserted into the Path message by the downstream PLR and MUST NOT be changed by downstream LSRs. The DOWNSTREAM_BYPASS_ASSIGNMENT object has the following format:

The IPv4 DOWNSTREAM_BYPASS_ASSIGNMENT object (Class-Num of the form 11bbbbbb with value = TBA, C-Type = TBA) has the following format:

| 0 | | | | | | | | | | 1 | | | | | | | | | | 2 | | | | | | | | | | 3 | | | | | | | | | |
|---------------------------------|---|---|---|---|---|---|---|---|---|-----------------|---|---|---|---|---|---|---|---|---|--------------|---|---|---|---|---|---|---|---|---|---|---|--|--|--|--|--|--|--|--|
| 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 0 | 1 | | | | | | | | |
| Length (bytes) | | | | | | | | | | Class-Num (TBA) | | | | | | | | | | C-Type (TBA) | | | | | | | | | | | | | | | | | | | |
| Bypass Tunnel ID | | | | | | | | | | Reserved | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| IPv4 Bypass Source Address | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| IPv4 Bypass Destination Address | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |

The IPv6 DOWNSTREAM_BYPASS_ASSIGNMENT object (Class-Num of the form 11bbbbbb with value = TBA, C-Type = TBA) has the following format:



Bypass Source Address

The bypass tunnel source IPV4 or IPV6 address.

Bypass Destination Address

The bypass tunnel destination IPV4 or IPV6 address.

Bypass Tunnel ID

The bypass tunnel identifier.

4.2. Bypass Tunnel Assignment Signaling Procedure

In cases where bidirectional bypass tunnels or a mix of unidirectional and bidirectional bypass tunnels are used for FRR Local Repair for a bidirectional co-routed LSP, it is desirable to coordinate the bypass tunnel selected at the downstream and upstream PLRs so that rerouted traffic and signaling flows on symmetrical paths post FRR. To achieve this, a new RSVP object is defined that identifies a bidirectional bypass tunnel that is assigned at a downstream PLR to protect a bidirectional LSP.

The DOWNSTREAM_BYPASS_ASSIGNMENT object is added by each downstream PLR in the RSVP Path message of the primary LSP to record the downstream bidirectional bypass tunnel assignment. This object is sent in the RSVP Path message every time the downstream PLR assigns or updates the bypass tunnel assignment so the upstream PLR may reflect the assignment too.

The upstream PLR (downstream MP) that detects a DOWNSTREAM_BYPASS_ASSIGNMENT object whose bypass tunnel destination matching its own address assigns the matching bidirectional bypass tunnel in the reverse direction, and removes the corresponding bypass tunnel assignment object before forwarding the RSVP Path message downstream. Otherwise, the bypass tunnel assignment object is forwarded downstream along in the RSVP Path message.

In absence of DOWNSTREAM_BYPASS_ASSIGNMENT object, the downstream MP can independently assign a bypass tunnel in the reverse direction. In the case of downstream MP receiving multiple DOWNSTREAM_BYPASS_ASSIGNMENT objects from multiple downstream PLRs, the decision of selecting a bypass tunnel in the reverse direction can be based on local policy, for example, prefer link protection vs. node protection bypass, or prefer the most upstream vs. least upstream node protection bypass tunnel. Note, the bypass tunnel selection will be corrected after FRR based on the PRR behavior after failure.

5. Compatibility

The DOWNSTREAM_BYPASS_ASSIGNMENT object to be defined with class numbers in the form 11bbbbbb, which ensures compatibility with non-supporting nodes. Per [RFC2205], nodes not supporting this extension will ignore the object but forward it, unexamined and unmodified, in all messages resulting from this message.

6. Security Considerations

This document introduces one new RSVP object. Thus in the event of

the interception of a signaling message, slightly more could be deduced about the state of the network than was previously the case, but this is judged to be a very minor security risk as this information is available by other means.

Otherwise, this document introduces no additional security considerations. For general discussion on MPLS and GMPLS related security issues, see the MPLS/GMPLS security framework [RFC5920].

7. IANA Considerations

A new Class-Num for the new DOWNSTREAM_BYPASS_ASSIGNMENT object is required.

8. Acknowledgements

Authors would like to thank George Swallow for his detailed and useful comments and suggestions.

9. References

9.1. Normative References

- [RSVP] Braden, R., Ed., Zhang, L., Berson, S., Herzog, S., and S. Jamin, "Resource ReSerVation Protocol (RSVP) -- Version 1 Functional Specification", RFC 2205, September 1997.
- [RSVP-TE] 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., Ed., Swallow, G., Ed., and A. Atlas, Ed., "Fast Reroute Extensions to RSVP-TE for LSP Tunnels", RFC 4090, May 2005.
- [RFC3473] Berger, L., Ed., "Generalized Multi-Protocol Label Switching (GMPLS) Signaling Resource ReserVation Protocol-Traffic Engineering (RSVP-TE) Extensions", RFC 3473, January 2003.
- [RFC3471] Berger, L., Ed., "Generalized Multi-Protocol Label Switching (GMPLS) Signaling Functional Description", RFC 3471, January 2003.

Authors' Addresses

Mike Taillon
Cisco Systems, Inc.
EMail: mtaillon@cisco.com

Tarek Saad
Cisco Systems, Inc.
EMail: tsaad@cisco.com

Rakesh Gandhi
Cisco Systems, Inc.
EMail: rgandhi@cisco.com

Zafar Ali
Cisco Systems, Inc.
EMail: zali@cisco.com

MPLS
Internet-Draft
Intended status: Informational
Expires: August 16, 2013

C. Villamizar, Ed.
OCCNC
K. Kompella
Contrail Systems
S. Amante
Level 3 Communications, Inc.
A. Malis
Verizon
C. Pignataro
Cisco
February 12, 2013

MPLS Forwarding Compliance and Performance Requirements
draft-villamizar-mpls-forwarding-01

Abstract

This document provides guidelines for implementors regarding MPLS forwarding and a basis for evaluations of forwarding implementations. Guidelines cover many aspects of MPLS forwarding. Topics are highlighted where implementors might potentially overlook practical requirements which are unstated or underemphasized or are optional for conformance to RFCs.

Status of this Memo

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

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

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

This Internet-Draft will expire on August 16, 2013.

Copyright Notice

Copyright (c) 2013 IETF Trust and the persons identified as the document authors. All rights reserved.

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

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

Table of Contents

| | |
|--|----|
| 1. Introduction | 4 |
| 1.1. Use of Requirements Language | 4 |
| 1.2. Apparent Misconceptions | 4 |
| 1.3. Target Audience | 5 |
| 2. Forwarding Issues | 6 |
| 2.1. Forwarding Basics | 6 |
| 2.1.1. MPLS Reserved Labels | 7 |
| 2.1.2. MPLS Differentiated Services | 7 |
| 2.1.3. Time Synchronization | 8 |
| 2.1.4. Uses of Multiple Label Stack Entries | 8 |
| 2.1.5. MPLS Link Bundling | 9 |
| 2.1.6. MPLS Hierarchy | 9 |
| 2.1.7. MPLS Fast Reroute (FRR) | 10 |
| 2.1.8. Pseudowire Encapsulation | 10 |
| 2.1.8.1. Pseudowire Sequence Number | 11 |
| 2.1.9. Layer-2 and Layer-3 VPN | 12 |
| 2.2. MPLS Multicast | 12 |
| 2.3. Packet Rates | 13 |
| 2.4. MPLS Multipath Techniques | 14 |
| 2.4.1. Pseudowire Control Word | 15 |
| 2.4.2. Large Microflows | 16 |
| 2.4.3. Pseudowire Flow Label | 16 |
| 2.4.4. MPLS Entropy Label | 16 |
| 2.4.5. Fields Used for Multipath | 17 |
| 2.4.5.1. MPLS Fields in Multipath | 17 |
| 2.4.5.2. IP Fields in Multipath | 19 |
| 2.4.5.3. Fields Used in Flow Label | 20 |
| 2.4.5.4. Fields Used in Entropy Label | 20 |
| 2.5. MPLS-TP and UHP | 21 |
| 2.6. OAM and DoS Protection | 21 |
| 2.6.1. DoS Protection | 21 |
| 2.6.2. MPLS OAM | 23 |
| 2.6.3. Pseudowire OAM | 24 |
| 2.6.4. MPLS-TP OAM | 25 |
| 2.6.5. MPLS OAM and Layer-2 OAM Interworking | 26 |
| 2.6.6. Extent of OAM Support by Hardware | 26 |

| | |
|--|----|
| 2.7. Number and Size of Flows | 27 |
| 3. Questions for Suppliers | 28 |
| 4. Forwarding Compliance and Performance Testing | 32 |
| 5. Acknowledgements | 37 |
| 6. IANA Considerations | 37 |
| 7. Security Considerations | 37 |
| 8. References | 38 |
| 8.1. Normative References | 38 |
| 8.2. Informative References | 39 |
| Appendix A. Organization of References Section | 43 |
| Authors' Addresses | 43 |

1. Introduction

The initial purpose of this document was to address concerns raised on the MPLS WG mailing list about shortcomings in implementations of MPLS forwarding. Documenting existing misconceptions and potential pitfalls might potentially avoid repeating past mistakes. The document has grown to address a broad set of forwarding requirements.

1.1. Use of Requirements Language

This document is informational. The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" are used only where the requirement is specified in an existing RFC. These keywords SHOULD be interpreted as described in RFC 2119 [RFC2119].

Advice given in this document does not use the upper case RFC 2119 keywords, except where explicitly noted that the keywords indicate that operator experiences indicate a requirement, but there are no existing RFC requirements. Such advice may be ignored by implementations. Similarly, implementations not claiming conformance to specific RFCs may ignore the requirements of those RFCs. In both cases, implementators may be doing so at their own peril.

1.2. Apparent Misconceptions

In early generations of forwarding silicon (which might now be behind us), there apparently were some misconceptions about MPLS. The following statements provide clarifications.

1. There are practical reasons to have more than one or two labels in an MPLS label stack. Under some circumstances the label stack can become quite deep. See Section 2.1.
2. The label stack MUST be considered to be arbitrarily deep. Section 3.27.4. "Hierarchy: LSP Tunnels within LSPs" of RFC 3031 [RFC3031] states "The label stack mechanism allows LSP tunneling to nest to any depth." If a the bottom of the label stack cannot be found, but sufficient number of labels exist to forward, an LSR MUST forward the packet. An LSR MUST NOT assume the packet is malformed unless the end of packet is found before bottom of stack. See Section 2.1.
3. In networks where deep label stacks are encountered, they are not rare. Full packet rate performance is required regardless of label stack depth, except where multiple POP operations are required. See Section 2.1.

4. Research has shown that long bursts of short packets with 40 byte or 44 byte IP payload sizes in these bursts are quite common. This is due to TCP ACK compression [ACK-compression].
 - A. A forwarding engine SHOULD, if practical, be able to sustain an arbitrarily long sequence of small packets arriving at full interface rate.
 - B. If indefinite full packet rate for small packets is not practical, a forwarding engine MUST be able to buffer a long sequence of small packets inbound to the on-chip decision engine and sustain full interface rate for some reasonable average packet rate. Absent this small on-chip buffering, QoS agnostic packet drops can occur.

See Section 2.3.

5. The implementor and system designer MUST support pseudowire control word if MPLS-TP is supported or if ACH is being used on a pseudowire [RFC5586]. Deployments SHOULD enable pseudowire control word. See Section 2.4.1.
6. The implementor and system designer SHOULD support adding a pseudowire Flow Label [RFC6391]. Deployments MAY enable this feature for appropriate pseudowire types. See Section 2.4.3.
7. The implementor and system designer SHOULD support adding a MPLS Entropy Label [RFC6790]. Deployments MAY enable this feature. See Section 2.4.4.

1.3. Target Audience

This document is intended for multiple audiences: implementor (implementing MPLS forwarding in silicon or in software); systems designer (putting together a MPLS forwarding systems); deployer (running an MPLS network). These guidelines are intended to serve the following purposes:

1. Explain what to do and what not to do when a deep label stack is encountered. (audience: implementor)
2. Highlight pitfalls to look for when implementing an MPLS forwarding chip. (audience: implementor)
3. Provide a checklist of features and performance specifications to request. (audience: systems designer, deployer)

4. Provide a set of tests to perform. (audience: systems designer, deployer).

The implementor, systems designer, and deployer have a transitive supplier customer relationship. It is in the best interest of the supplier to review their product against their customer's checklist and customer's customer's checklist if applicable.

2. Forwarding Issues

A brief review of forwarding issues is provided in the subsections that follow. This section provides some background on why some of these requirements exist. The questions to ask of suppliers and testing is covered in the following sections, Section 3 and Section 4.

2.1. Forwarding Basics

Basic MPLS architecture and MPLS encapsulation, and therefore packet forwarding is defined in [RFC3031] and [RFC3032]. RFC3031 and RFC3032 are somewhat LDP centric. RSVP-TE supports traffic engineering (TE) and fast reroute, features that LDP lacks. The base document for RSVP-TE based MPLS is [RFC3209].

A few RFCs update RFC3032. Those with impact on forwarding include the following.

1. TTL processing is clarified in [RFC3443].
2. The use of MPLS Explicit NULL is modified in [RFC4182].
3. Differentiated Services is supported by [RFC3270] and [RFC4124]. The "EXP" field is renamed to "Traffic Class" in [RFC5462], removing any misconception that it was available for experimentation or could be ignored.
4. ECN is supported by [RFC5129].
5. The MPLS G-ACh and GAL are defined in [RFC5586].

Other RFCs have implications to MPLS Forwarding and do not update RFC3032 or RFC3209, including:

1. The pseudowire (PW) Associated Channel Header (ACH), defined by [RFC5085], later generalized by the MPLS G-ACh [RFC5586].

2. The Entropy Label Indicator and Entropy Label are defined by [RFC6790].

A few RFCs update RFC3209. Those that are listed as updating RFC3209 generally impact only RSVP-TE signaling. Forwarding is modified by major extension built upon RFC3209.

RFCs which impact forwarding are discussed in the following subsections.

2.1.1. MPLS Reserved Labels

[RFC3032] specifies that label values 0-15 are reserved labels with special meanings. Three values of NULL label are defined (two of which are later updated by [RFC4182]) and a router-alert label is defined. The original intent was that reserved labels, except the NULL labels, could be sent to the routing engine CPU rather than be processed in forwarding hardware. Hardware support is required by new RFCs such as those defining Entropy Label and OAM processed as a result of receiving a GAL. For new reserved labels, some accommodation is needed for LSR that will send the labels to a general purpose CPU. For example, ELI will only be sent to LSR which have signaled support for [RFC6790] and high OAM packet rate must be negotiated among endpoints.

[RFC3429] reserves a label for ITU-T Y.1711, however Y.1711 does not work with multipath and its use is strongly discouraged.

The current list of reserved labels can be found on the "Multiprotocol Label Switching Architecture (MPLS) Label Values" registry reachable at IANA's pages at <<http://www.iana.org>>.

When an unknown reserved label is encountered or a reserved label not directly handled in forwarding hardware is encountered, the packet should be sent to a general purpose CPU by default. If this capability is supported, there must be an option to either drop or rate limit such packets on a per reserved label value basis.

2.1.2. MPLS Differentiated Services

[RFC2474] deprecates the IP Type of Service (TOS) and IP Precedence (Prec) fields and replaces them with the Differentiated Services Field more commonly known as the Differentiated Services Code Point (DSCP) field. [RFC2475] defines the Differentiated Services architecture, which in other forum is often called a Quality of Service (QoS) architecture.

MPLS uses the Traffic Class (TC) field to support Differentiated

Services [RFC5462]. There are two primary documents describing how DSCP is mapped into TC.

1. [RFC3270] defines E-LSP and L-LSP. E-LSP use a static mapping of DSCP into TC. L-LSP use a per LSP mapping of DSCP into TC, with one PHB Scheduling Class (PSC) per L-LSP. Each PSC can use multiple Per-Hop Behavior (PHB) values. For example, the Assured Forwarding service defines three PSC, each with three PHB [RFC2597].
2. [RFC4124] defines assignment of a class-type (CT) to an LSP, where a per CT static mapping of TC to PHB is used. [RFC4124] provides a means to support up to eight E-LSP-like mappings of DSCP to TC.

To meet Differentiated Services requirements specified in [RFC3270], the following forwarding requirements must be met. An ingress LER MUST be able to select an LSP and then apply a per LSP map of DSCP into TC. A midpoint LSR MUST be able to apply a per LSP map of TC to PHB. The number of mappings supported will be far less than the number of LSP supported.

2.1.3. Time Synchronization

PTP or NTP may be carried over MPLS [I-D.ietf-tictoc-1588overmpls]. Generally NTP will be carried within IP with IP carried in MPLS [RFC5905]. Both PTP and NTP benefit from accurate time stamping of incoming packets and the ability to insert accurate time stamps in outgoing packets.

Since the label stack depth may vary, hardware should allow a timestamp to be placed in an outgoing packet at any specified byte position. It may be necessary to modify layer-2 checksums or frame check sequences after insertion. PTP and NTP timestamp formats differ slightly.

Accurate time synchronization in addition to being generally useful is required for MPLS-TP delay measurement (DM) OAM. See Section 2.6.4.

2.1.4. Uses of Multiple Label Stack Entries

MPLS deployments in the early part of the prior decade (circa 2000) tended to support either LDP or RSVP-TE. LDP was favored by some for its ability to scale to a very large number of PE devices at the edge of the network, without adding deployment complexity. RSVP-TE was favored, generally in the network core, where traffic engineering and/or fast reroute were considered important.

Both LDP and RSVP-TE are used simultaneously within major Service Provider networks using a technique known as "LDP over RSVP-TE Tunneling". This technique allows service providers to carry LDP tunnels, originating and terminating at PE's, inside of RSVP-TE tunnels, generally between Inter-City P routers, to take advantage of Traffic Engineering and Fast Re-Route on more expensive Inter-City and Inter-Continental Transport paths. LDP over RSVP-TE tunneling requires a minimum of two MPLS labels: one each for LDP and RSVP-TE.

The use of MPLS FRR [RFC4090] added one more label to MPLS traffic, but only when FRR protection was in use. If LDP over RSVP-TE is in use, and FRR protection is in use, then at least three MPLS labels are present on the label stack on the links through which the Bypass LSP traverses. FRR is covered in Section 2.1.7.

LDP L2VPN, LDP IPVPN, BGP L2VPN, and BGP IPVPN added support for VPN services that are deployed in the vast majority of service providers. These VPN services added yet another label, bringing the label stack depth (when FRR is active) to four.

Pseudowires and VPN are discussed in further detail in Section 2.1.8 and Section 2.1.9.

2.1.5. MPLS Link Bundling

MPLS Link Bundling was the first RFC to address the need for multiple parallel links between nodes [RFC4201]. MPLS Link Bundling is notable in that it tried not to change MPLS forwarding, except in specifying the "All-Ones" component link. MPLS Link Bundling is seldom if ever deployed. Instead multipath techniques described in Section 2.4 are used.

2.1.6. MPLS Hierarchy

MPLS hierarchy is defined in [RFC4206]. Although RFC4206 is considered part of GMPLS, the Packet Switching Capable (PSC) portion of the MPLS hierarchy are applicable to MPLS and may be supported in an otherwise GMPLS free implementation. The MPLS PSC hierarchy remains the most likely means of providing further scaling in an RSVP-TE MPLS network, particularly where the network is designed to provide RSVP-TE connectivity to the edges. This is the case for envisioned MPLS-TP networks. The use of the MPLS PSC hierarchy can add as many as four labels to a label stack, though it is likely that only one layer of PSC will be used in the near future.

2.1.7. MPLS Fast Reroute (FRR)

Fast reroute is defined by [RFC4090]. Two significantly different methods are the "One-to-One Backup" method which uses the "Detour LSP" and the "Facility Backup" which uses a "bypass tunnel". These are commonly referred to as the detour and bypass methods respectively.

The detour method makes use of a presignaled LSP. Hardware assistance is needed for detour FRR only if necessary to accomplish local repair of a large number of LSP within the 10s of milliseconds target. For each affected LSP a SWAP operation must be reprogrammed or otherwise switched over. The use of detour FRR doubles the number of LSP terminating at any given hop and will increase the number of LSP within a network by a factor dependent on the average detour path length.

The bypass method makes use of a tunnel that is unused when no fault exists but may carry many LSP when a local repair is required. There is no presignaling indicating which working LSP will be diverted into any specific bypass LSP. The egress LSR of the bypass LSP MUST use platform label space (as defined in [RFC3031]) so that an LSP working path on any give interface can be backed up using a bypass LSP terminating on any other interface. Hardware assistance is needed if necessary to accomplish local repair of a large number of LSP within the 10s of milliseconds target. For each affected LSP a SWAP operation must be reprogrammed or otherwise switched over with an additional PUSH of the bypass LSP label. In addition the use of platform label space impacts the size of the LSR ILM for LSR with a very large number of interfaces.

2.1.8. Pseudowire Encapsulation

The pseudowire (PW) architecture is defined in [RFC3985]. A pseudowire, when carried over MPLS, adds one or more additional label entries to the MPLS label stack. A PW Control Word is defined in [RFC4385] with motivation for defining the control word in [RFC4928]. The PW Associated Channel defined in [RFC4385] is used for OAM in [RFC5085]. The PW Flow Label is defined in [RFC6391] and is discussed further in this document in Section 2.4.3.

There are numerous pseudowire encapsulations, supporting emulation of services such as Frame Relay, ATM, Ethernet, TDM, and SONET/SDH over packet switched networks (PSNs) using IP or MPLS.

The pseudowire encapsulation is out of scope for this document. Pseudowire impact on MPLS forwarding at midpoint LSR is within scope. The impact on ingress MPLS PUSH and egress MPLS UHP POP are within

scope. While pseudowire encapsulation is out of scope, some advice is given on sequence number support.

2.1.8.1. Pseudowire Sequence Number

Pseudowire (PW) sequence number support is most important for PW payload types with a high expectation of in-order delivery. Resequencing support, rather than dropping at egress on out of order arrival, is most important for PW payload types with a high expectation of lossless delivery. For example, TDM payloads require sequence number support and require resequencing support. The same is true of ATM CBR service. ATM VBR or ABR may have somewhat relaxed requirements, but generally require ATM Early Packet Discard (EPD) or ATM Partial Packet Discard (PPD) [ATM-EPD-and-PPD]. Though sequence number support and resequencing support are beneficial to PW packet oriented payloads such as FR and Ethernet, they are highly desirable but not as strongly required.

Packet reorder should be rare except in a small number of circumstances, most of which are due to network design or equipment design errors:

1. The most common case is where reordering occurs is rare, occurring only when a network or equipment fault forces traffic on a new path with different delay. The packet loss that accompanies a network or equipment fault is generally more disruptive than any reordering which may occur.
2. A path change can be caused by reasons other than a network or equipment fault, such as administrative routing change. This may result in packet reordering but generally without any packet loss.
3. If the edge is not using pseudowire control word (CW) and the core is using multipath, reordering will be far more common. If this is occurring, the best solution is to use CW on the edge, rather than try to fix the reordering using resequencing.
4. Another avoidable case is where some core equipment has multipath and for some reason insists on periodically installing a new random number as the multipath hash seed. If supporting MPLS-TP, equipment MUST provide a means to disable periodic hash reseeding and deployments MUST disable periodic hash reseeding. Even if not supporting MPLS-TP, equipment should provide a means to disable periodic hash reseeding and deployments should disable periodic hash reseeding.

2.1.9. Layer-2 and Layer-3 VPN

Layer-2 VPN [RFC4664] and Layer-3 VPN [RFC4110] add one or more label entry to the MPLS label stack. VPN encapsulations are out of scope for this document. Its impact on forwarding at midpoint LSR are within scope.

Any of these services may be used on an MPLS Entropy Label enabled ingress and egress (see Section 2.4.4 for discussion of Entropy Label) which would add an additional label to the MPLS label stack. The need to provide a useful Entropy Label value impacts the requirements of the VPN ingress LER but is out of scope for this document.

2.2. MPLS Multicast

MPLS Multicast encapsulation is clarified in [RFC5332]. MPLS Multicast may be signaled using RSVP-TE [RFC4875] or LDP [RFC6388].

[RFC4875] defines a root initiated RSVP-TE LSP setup rather than leaf initiated join used in IP multicast. [RFC6388] defines a leaf initiated LDP setup. Both [RFC4875] and [RFC6388] define point to multipoint (P2MP) LSP setup. [RFC6388] also defined multipoint to multipoint (MP2MP) LSP setup.

The P2MP LSP have a single source. An LSR may be a leaf node, an intermediate node, or a "bud" node. A bud serves as both a leaf and intermediate. At a leaf an MPLS POP is performed. The payload may be a IP Multicast packet that requires further replication. At an intermediate node a MPLS SWAP is performed. The bud requires that both a POP and SWAP be performed for the same incoming packet.

One strategy to support P2MP functionality is to POP at the LSR ingress and then optionally PUSH labels at each LSR egress. A given LSR egress chip may support multiple egress interfaces, each of which requires a copy, but each with a different set of added labels and layer-2 encapsulation. Some physical interfaces may have multiple sub-interfaces (such as Ethernet VLAN or channelized interfaces) each requiring a copy.

If packet replication is performed at LSR ingress, then the ingress interface performance may suffer. If the packet replication is performed within a LSR switching fabric and at LSR egress, congestion of egress interfaces cannot make use of backpressure to ingress interfaces using techniques such as virtual output queuing (VOQ). If buffering is primarily supported at egress, then the need for backpressure is minimized. There may be no good solution for high volumes of multicast traffic if VOQ is used.

MP2MP LSP differ in that any branch may provide an input, including a leaf. Packets must be replicated onto all other branches. This forwarding is often implemented as multiple P2MP forwarding trees, one for each potential input.

2.3. Packet Rates

While average packet size of Internet traffic may be large, long sequences of small packets have both been predicted in theory and observed in practice. Traffic compression and TCP ACK compression can conspire to create long sequences of packets of 40-44 bytes in payload length. If carried over Ethernet, the 64 byte minimum payload applies, yielding a packet rate of approximately 150 Mpps (million packets per second) for the duration of the burst on a nominal 100 Gb/s link. The peak rate is higher for other encapsulations can be as high as 250 Mpps (for example IP or MPLS encapsulated using GFP over OTN ODU4).

It is also possible that the packet rates for a minimum payload size, such as 64 byte (64B) payload for Ethernet, is acceptable, but the rate declines for other packet sizes, such as 65B payload. There are other packet rates of interest besides TCP ACK. For example, a TCP ACK carried over an Ethernet PW over MPLS over Ethernet may occupy 82B or 82B plus an increment of 4B if additional MPLS labels are present.

A graph of packet rate vs. packet size often displays a sawtooth. The sawtooth is commonly due to a memory bottleneck and memory widths, sometimes internal cache, but often a very wide external buffer memory interface. In some cases it may be due to a fabric transfer width. A fine packing, rounding up to the nearest 8B or 16B will result in a fine sawtooth with small degradation for 65B, and even less for 82B packets. A coarse packing, rounding up to 64B can yield a sharper drop in performance for 65B packets, or perhaps more important, a larger drop for 82B packets.

The loss of some TCP ACK packets are not the primary concern when such a burst occurs. When a burst occurs, any other packets, regardless of packet length and packet QoS are dropped once on-chip input buffers prior to the decision engine are exceeded. Buffers in front of the packet decision engine are often very small or non-existent (less than one packet of buffer) causing significant QoS agnostic packet drop.

Internet service providers and content providers generally specify full rate forwarding with 40 byte payload packets as a requirement. This requirement often can be waived if the provider can be convinced that when long sequence of short packets occur no packets will be

dropped.

Many equipment suppliers have pointed out that the extra cost in designing hardware capable of processing the minimum size packets at full line rate is significant for very high speed interfaces. If hardware is not capable of processing the minimum size packets at full line rate, then that hardware MUST be capable of handling large burst of small packets, a condition which is often observed. This level of performance is necessary to meet Differentiated Services [RFC2475] requirements for without it, packets are lost prior to inspection of the IP DSCP field [RFC2474] or MPLS TC field [RFC5462].

With adequate on-chip buffers before the packet decision engine, an LSR can absorb a long sequence of short packets. Even if the output is slowed to the point where light congestion occurs, the packets, having cleared the decision process, can make use of larger VOQ or output side buffers and be dealt with according to configured QoS treatment, rather than dropped completely at random.

These on-chip buffers need not contribute significant delay since they are only used when the packet decision engine is unable to keep up, not in response to congestion, plus these buffers are quite small. For example, an on-chip buffer capable of handling 4K packets of 64 bytes in length, or 256KB, corresponds to 2 msec on a 10 Mb/s link and 0.2 usec on a 100 Gb/s link. If the packet decision engine is capable of handling packets at 90% of the full rate for small packets, then the maximum added delay is 0.2 msec and 20 nsec respectively, and this delay only applies if a 4K burst of short packets occurs. When no burst of short packets was being processed, no delay is added.

Packet rate requirements apply regardless of which network tier equipment is deployed in. Whether deployed in the network core or near the network edges, one of the two conditions MUST be met:

1. Packets must be processed at full line rate with minimum sized packets. -OR-
2. Packets must be processed at a rate well under generally accepted average packet sizes, with sufficient buffering prior to the packet decision engine to accommodate long bursts of small packets.

2.4. MPLS Multipath Techniques

In any large provider, service providers and content providers, hash based multipath techniques are used in the core and in the edge. In many of these providers hash based multipath is also used in the

larger metro networks.

The most common multipath techniques are ECMP applied at the IP forwarding level, Ethernet LAG with inspection of the IP payload, and multipath on links carrying both IP and MPLS, where the IP header is inspected below the MPLS label stack. In most core networks, the vast majority of traffic is MPLS encapsulated.

In order to support an adequately balanced load distribution across multiple links, IP header information must be used. Common practice today is to reinspect the IP headers at each LSR and use the label stack and IP header information in a hash performed at each LSR. Further details are provided in Section 2.4.5.

The use of this technique is so ubiquitous in provider networks that lack of support for multipath makes any product unsuitable for use in large core networks. This will continue to be the case in the near future, even as deployment of MPLS Entropy Label begins to relax the core LSR multipath performance requirements given the existing deployed base of edge equipment without the ability to add an Entropy Label.

A generation of edge equipment supporting the ability to add an MPLS Entropy Label is needed before the performance requirements for core LSR can be relaxed. However, it is likely that two generations of deployment in the future will allow core LSR to support full packet rate only when a relatively small number of MPLS labels need to be inspected before hashing. For now, don't count on it.

Common practice today is to reinspect the packet at each LSR and use the label stack and use the IP header field as input to a hash algorithm performed on each packet at each LSR in the network combined with a hash seed that is selected by each LSR. Where flow labels or entropy labels are used, a hash seed must be used.

2.4.1. Pseudowire Control Word

Within the core of a network some form of multipath is almost certain to be used. Multipath techniques deployed today are likely to be looking beneath the label stack for an opportunity to hash on IP addresses.

A pseudowire encapsulated at a network edge must have a means to prevent reordering within the core if the pseudowire will be crossing a network core, or any part of a network topology where multipath is used (see [RFC4385] and [RFC4928]).

Not supporting the ability to encapsulate a pseudowire with a control

word may lock a product out from consideration. A pseudowire capability without control word support might be sufficient for applications that are strictly both intra-metro and low bandwidth. However a provider with other applications will very likely not tolerate having equipment which can only support a subset of their pseudowire needs.

2.4.2. Large Microflows

Where multipath makes use of a simple hash and simple load balance such as modulo or other fixed allocation (see Section 2.4) the presence of large microflows that each consumes 10% of the capacity of a component link of a potentially congested composite link, one such microflow can upset the traffic balance and more than one can in effect reduce the effective capacity of the entire composite link by more than 10%.

When even a very small number of large microflows are present, there is a significant probability that more than one of these large microflows could fall on the same component link. If the traffic contribution from large microflows is small, the probability for three or more large microflows on the same component link drops significantly. Therefore in a network where a significant number of parallel 10 Gb/s links exists, even a 1 Gb/s pseudowire or other large microflow that could not otherwise be subdivided into smaller flows should carry a flow label or entropy label if possible.

Active management of the hash space to better accommodate large microflows has been implemented and deployed in the past, however such techniques are out of scope for this document.

2.4.3. Pseudowire Flow Label

Unlike a pseudowire control word, a pseudowire flow label [RFC6391], is required only for relatively large capacity pseudowires. There are many cases where a pseudowire flow label makes sense. Any service such as a VPN which carries IP traffic within a pseudowire can make use of a pseudowire flow label.

Any pseudowire carried over MPLS which makes use of the pseudowire control word and does not carry a flow label is in effect a single microflow (in [RFC2475] terms).

2.4.4. MPLS Entropy Label

The MPLS Entropy Label simplifies flow group identification [RFC6790] in the network core. Prior to the MPLS Entropy Label core LSR needed to inspect the entire label stack and often the IP headers to provide

an adequate distribution of traffic when using multipath techniques (see Section 2.4.5). With the use of MPLS Entropy Label, a hash can be performed closer to network edges, placed in the label stack, and used within the network core.

The MPLS Entropy Label is capable of avoiding full label stack and payload inspection within the core where performance levels are most difficult to achieve (see Section 2.3). The label stack inspection can be terminated as soon as the first Entropy Label is encountered, which is generally after a small number of labels are inspected.

In order to provide these benefits in the core, LSR closer to the edge must be capable of adding an entropy label. This support may not be required in the access tier, the tier closest to the customer, but is likely to be required in the edge or the border to the network core. LSR peering with external networks will also need to be able to add an Entropy Label.

2.4.5. Fields Used for Multipath

The most common multipath techniques are based on a hash over a set of fields. Regardless of whether a hash is used or some other method is used, there are a limited set of fields which can safely be used for multipath.

2.4.5.1. MPLS Fields in Multipath

If the "outer" or "first" layer of encapsulation is MPLS, then label stack entries are used in the hash. Within a finite amount of time (and for small packets arriving at high speed that time can quite limited) only a finite number of label entries can be inspected. Pipelined or parallel architectures improve this, but the limit is still finite.

The following guidelines are provided for use of MPLS fields in multipath load balancing.

1. Only the 20 bit label field SHOULD be used. The TTL field SHOULD NOT be used. The S bit MUST NOT be used. The TC field (formerly EXP) MUST NOT be used. See below this list for reasons.
2. If an ELI label is found, then if the LSR supports Entropy Label, the EL label field in the next label entry (the EL) SHOULD be used and label entries below that label SHOULD NOT be used and the MPLS payload SHOULD NOT be used. See below this list for reasons.

3. Reserved labels (label values 0-15) MUST NOT be used. In particular, GAL and RA MUST NOT be used so that OAM traffic follows the same path as payload packets with the same label stack.
4. The most entropy is generally found in the label stack entries near the bottom of the label stack (innermost label, closest to S=1 bit). If the entire label stack cannot be used (or entire stack up to an EL), then it is better to use as many labels as possible closest to the bottom of stack.
5. If no ELI is encountered, and the first nibble of payload contains a 4 (IPv4) or 6 (IPv6), an implementation SHOULD support the ability to interpret the payload as IPv4 or IPv6 and extract and use appropriate fields from the IP headers. This feature is considered a hard requirement by many service providers. If supported, there MUST be a way to disable it (if, for example, PW without CW are used). This ability to disable this feature is considered a hard requirement by many service providers. Therefore an implementation has a very strong incentive to support both options.
6. A label which is popped at egress (UHP POP) SHOULD NOT be used. A label which is popped at the penultimate hop (PHP POP) SHOULD be used.

Apparently some chips have made use of the TC (formerly EXP) bits as a source of entropy. This is very harmful since it will reorder Assured Forwarding (AF) traffic [RFC2597] when a subset does not conform to the configured rates and is remarked but not dropped at a prior LSR. Traffic which uses MPLS ECN [RFC5129] can also be reordered if TC is used for entropy. Therefore, as stated in the guidelines above, the TC field (formerly EXP) MUST NOT be used in multipath load balancing as it violates Differentiated Services Ordered Aggregate (OA) requirements in these two instances.

Use of the MPLS label entry S bit would result in putting OAM traffic on a different path if the addition of a GAL at the bottom of stack removed the S bit from the prior label.

If an ELI label is found, then if the LSR supports Entropy Label, the EL label field in the next label entry (the EL) SHOULD be used and the search for additional entropy within the packet SHOULD be terminated. Failure to terminate the search will impact client MPLS-TP LSP carried within server MPLS LSP. A network operator has the option to use administrative attributes as a means to identify LSR which do not terminate the entropy search at the first EL. Administrative attributes are defined in [RFC3209]. Some

configuration is required to support this.

If the PHP POP label is not used, then for any PW for which CW is used, there is no basis for multipath load split. In some networks it is infeasible to put all PW traffic on one component link. Any PW which does not use CW will be improperly split regardless of whether the PHP POP label is used.

2.4.5.2. IP Fields in Multipath

Inspecting the IP payload provides the most entropy in provider networks. The practice of looking past the bottom of stack label for an IP payload is well accepted and documented in [RFC4928] and in other RFCs.

Where IP is mentioned in the document, both IPv4 and IPv6 apply. All LSRs MUST fully support IPv6.

When information in the IP header is used, the following guidelines apply:

1. Both the IP source address and IP destination address SHOULD be used. There MAY be an option to reverse the order of these address, improving the ability to provide symmetric paths in some cases. Many service providers require that both addresses be used.
2. Implementations SHOULD allow inspection of the IP protocol field and use of the UDP or TCP port numbers. For many service providers this feature is considered mandatory, particularly for enterprise, data center, or edge equipment. If this feature is provided, it SHOULD be possible to disable use of TCP and UDP ports. Many service providers consider it a hard requirement that use of UDP and TCP ports can be disabled. Therefore there is a strong incentive for implementations to provide both options.
3. Equipment suppliers MUST NOT make assumptions that because the IP version field is equal to 4 (an IPv4 packet) that the IP protocol will either be TCP (IP protocol 6) or UDP (IP protocol 17) and blindly fetch the data at the offset where the TCP or UDP ports would be found. With IPv6, TCP and UDP port numbers are not at fixed offsets. With IPv4 packets carrying IP options, TCP and UDP port numbers are not at fixed offsets.
4. The IPv6 header flow field SHOULD be used. This is the explicit purpose of the IPv6 flow field, however observed flow fields rarely contains a non-zero value. Some uses of the flow field have been defined such as [RFC6438]. In the absence of MPLS

encapsulation, the IPv6 flow field can serve a role equivalent to Entropy Label.

5. Support other protocols that share a common Layer-4 header such as RTP, UDP-lite, SCTP and DCCP SHOULD be provided, particularly for edge or access equipment where additional entropy may be needed. Equipment SHOULD also use RTP, UDP-lite, SCTP and DCCP headers when creating an Entropy Label.
6. Similar to avoiding TC in MPLS, the IP DSCP, and ECN bits MUST NOT be used. The IPv4 TTL or IPv6 Hop Count SHOULD NOT be used. Note that the IP TOS field was deprecated ([RFC0791] was updated by [RFC2474]). No part of the IP DSCP (formerly IP PREC and IP TOS bits) field can be used.
7. Some IP encapsulations support tunneling, such as IP-in-IP, GRE, L2TPv3, and IPSEC. These provide a greater source of entropy which some provider networks carrying large amounts of tunneled traffic may need. The use of tunneling header information is out of scope for this document.

This document makes the following recommendations. These recommendations are not required to claim compliance to any existing RFC therefore implementors are free to ignore them, but due to service provider requirements may be doing so at their own peril. The use of IP addresses MUST be supported and TCP and UDP ports (conditional on the protocol field and properly located) MUST be supported. The ability to disable use of UDP and TCP ports MUST be available. Though potentially very useful in some networks, it is uncommon to support using payloads of tunneling protocols carried over IP. Though the use of tunneling protocol header information is out of scope for this document, it is not discouraged.

2.4.5.3. Fields Used in Flow Label

The ingress to a pseudowire (PW) can extract information from the payload being encapsulated to create a flow label. [RFC6391] references IP carried in Ethernet as an example. The Native Service Processing (NSP) function defined in [RFC3985] differs with pseudowire type. It is in the NSP function where information for a specific type of PW can be extracted for use in a flow label. Which fields to use for any given PW NSP is out of scope for this document.

2.4.5.4. Fields Used in Entropy Label

An entropy label is added at the ingress to an LSP. The payload being encapsulated is most often MPLS, a PW, or IP. The payload type is identified by the layer-2 encapsulation (Ethernet, GFP, POS, etc).

If the payload is MPLS, then the information used to create an entropy label is the same information used for local load balancing (see Section 2.4.5.1). This information **MUST** be extracted for use in generating an entropy label even if the LSR local egress interface is not a multipath.

Of the non-MPLS payload types, only payloads that are forwarded are of interest. For example, ARP is not forwarded and CNLP (used only for ISIS) is not forwarded.

The non-MPLS payload type of greatest interest are IPv4 and IPv6. The guidelines in Section 2.4.5.2 apply to fields used to create and entropy label.

The IP tunneling protocols mentioned in Section 2.4.5.2 may be more applicable to generation of an entropy label at edge or access where deep packet inspection is practical due to lower interface speeds than in the core where deep packet inspection may be impractical.

2.5. MPLS-TP and UHP

MPLS-TP introduces forwarding demands that will be extremely difficult to meet in a core network. Most troublesome is the requirement for Ultimate Hop Popping (UHP, the opposite of Penultimate Hop Popping or PHP). Using UHP opens the possibility of one or more MPLS POP operation plus an MPLS SWAP operation for each packet. The potential for multiple lookups and multiple counter instances per packet exists.

As networks grow and tunneling of LDP LSPs into RSVP-TE LSPs is used, and/or RSVP-TE hierarchy is used, the requirement to perform one or two or more MPLS POP operations plus a MPLS SWAP operation (and possibly a PUSH or two) increases. If MPLS-TP LM (link monitoring) OAM is enabled at each layer, then a packet and byte count **MUST** be maintained for each POP and SWAP operation so as to offer OAM for each layer.

2.6. OAM and DoS Protection

Denial of service (DoS) protection is an area requiring hardware support that is often overlooked or inadequately considered. Hardware assist is also needed for OAM, particularly the more demanding MPLS-TP OAM.

2.6.1. DoS Protection

Modern equipment supports a number of control plane and management plane protocols. Generally no single means of protecting network

equipment from denial of service (DoS) attacks is sufficient, particularly for high speed interfaces. This problem is not specific to MPLS, but is a topic that cannot be ignored when implementing or evaluating MPLS implementations.

Two types of protections are often cited as primary means of protecting against attacks of all kinds.

Isolated Control/Management Traffic

Control and Management traffic can be carried out-of-band (OOB), meaning not intermixed with payload. For MPLS use of G-ACh and GAL to carry control and management traffic provides a means of isolation from potentially malicious payload. Used along, the compromise of a single node, including a small computer at a network operations center, could compromise an entire network. Implementations which send all G-ACh/GAL traffic directly to a routing engine CPU are subject to DoS attack as a result of such a compromise.

Cryptographic Authentication

Cryptographic authentication can very effectively prevent malicious injection of control or management traffic. Cryptographic authentication can in some circumstances be subject to DoS attack by overwhelming the capacity of the decryption with a high volume of malicious traffic. For very low speed interfaces cryptographic authentication can be performed by the general purpose CPU used as a routing engine. For all other cases, cryptographic hardware may be needed. For very high speed interfaces, even cryptographic hardware can be overwhelmed.

Some control and management protocols are often carried with payload traffic. This is commonly the case with BGP, T-LDP, and SNMP. It is often the case with RSVP-TE. Even when carried over G-ACh/GAL additional measures can reduce the potential for a minor breach to be leveraged to a full network attack.

Some of the additional protections are supported by hardware packet filtering.

GTSM

[RFC5082] defines a mechanism that uses the IPv4 TTL or IPv6 Hop Limit fields to insure control traffic that can only originate from an immediate neighbor is not forged and originating from a distant source. GTSM can be applied to many control protocols which are routable, for example LDP [RFC6720].

IP Filtering

At the very minimum, packet filtering plus classification and use of multiple queues supporting rate limiting is needed for traffic that could potentially be sent to a general purpose CPU used as a routing engine. The first level of filtering only allows connections to be initiated from specific IP prefixes to specific destination ports and then preferably passes traffic directly to a cryptographic engine and/or rate limits. The second level of filtering passes connected traffic, such as TCP connections having received at least one authenticated SYN or having been locally initiated. The second level of filtering only passes traffic to specific address and port pairs to be checked for cryptographic authentication.

The cryptographic authentication is generally the last resort in DoS attack mitigation. If a packet must be first sent to a general purpose CPU, then sent to a cryptographic engine, a DoS attack is possible on high speed interfaces. Only where hardware can identify a signature and the portion of packet covered by the signature is cryptographic authentication highly beneficial in protecting against DoS attacks.

For chips supporting multiple 100 Gb/s interfaces, only a very large number of parallel cryptographic engines can provide the processing capacity to handle a large scale DoS or distributed DoS (DDoS) attack. For many forwarding chips this much processing power requires significant chip real estate and power, and therefore reduces system space and power density. For this reason, cryptographic authentication is not considered a viable first line of defense.

For some networks the first line of defense is some means of supporting OOB control and management traffic. In the past this OOB channel might make use of overhead bits in SONET or OTN or a dedicated DWDM wavelength. G-ACh and GAL provide an alternative OOB mechanism which is independent of underlying layers. In other networks, including most IP/MPLS networks, perimeter filtering serves a similar purpose, though less effective without extreme vigilance.

A second line of defense is filtering, including GTSM. For protocols such as EBGp, GTSM and other filtering is often the first line of defense. Cryptographic authentication is usually the last line of defense and insufficient by itself to mitigate DoS or DDoS attacks.

2.6.2. MPLS OAM

[RFC4377] defines requirements for MPLS OAM that predate MPLS-TP.
[RFC4379] defines what is commonly referred to as LSP Ping and LSP

Traceroute. [RFC4379] is updated by [RFC6424] supporting MPLS tunnels and stitched LSP and P2MP LSP. [RFC4379] is updated by [RFC6425] supporting P2MP LSP. [RFC4379] is updated by [RFC6426] to support MPLS-TP connectivity verification (CV) and route tracing.

[RFC4950] extends the ICMP format to support TTL expiration that may occur when using IP traceroute within an MPLS tunnel. The ICMP message generation can be implemented in forwarding hardware, but if sent to a general purpose CPU must be rate limited to avoid a potential denial or service (DoS) attack.

[RFC5880] defines Bidirectional Forwarding Detection (BFD), a protocol intended to detect faults in the bidirectional path between two forwarding engines. [RFC5884] and [RFC5885] define BFD for MPLS. BFD can provide failure detection on any kind of path between systems, including direct physical links, virtual circuits, tunnels, MPLS Label Switched Paths (LSPs), multihop routed paths, and unidirectional links as long as there is some return path.

The processing requirements for BFD are less than for LSP Ping, making BFD somewhat better suited for relatively high rate proactive monitoring. BFD does not verify that the data plane against the control plane, where LSP Ping does. LSP Ping somewhat better suited for on-demand monitoring including relatively low rate periodic verification of data plane and as a diagnostic tool.

Both BFD and LSP Ping MUST be recognized by hardware and at the very minimum forwarded to the main CPU. Hardware assistance for BFD is often provided and is considered necessary for relatively high rate proactive monitoring. Both BFD and LSP Ping MUST be recognized in any filtering prior to passing traffic to a general purpose CPU and appropriate DoS protection applied (see Section 2.6.1. Failure to recognize BFD and LSP Ping and at least rate limit creates the potential for misconfiguration to cause outages rather than cause errors in the misconfigured OAM.

2.6.3. Pseudowire OAM

Pseudowire OAM makes use of the control channel provided by Virtual Circuit Connectivity Verification (VCCV) [RFC5085]. VCCV makes use of the Pseudowire Control Word. BFD support over VCCV is defined by [RFC5885]. [RFC5885] is updated by [RFC6478] in support of static pseudowires. [RFC4379] is updated by [RFC6829] supporting LSP Ping for Pseudowire FEC advertised over IPv6.

G-ACh/GAL (defined in [RFC5586]) is the preferred MPLS-TP OAM control channel and applies to any MPLS-TP end points, including Pseudowire. See Section 2.6.4 for an overview of MPLS-TP OAM.

2.6.4. MPLS-TP OAM

[RFC6669] summarizes the MPLS-TP OAM toolset, the set of protocols supporting the MPLS-TP OAM requirements specified in [RFC5860] and supported by the MPLS-TP OAM framework defined in [RFC6371].

The MPLS-TP OAM toolset includes:

CC-CV

[RFC6428] defines BFD extensions to support proactive CC-CV applications. [RFC6426] provides LSP ping extensions that are used to implement on-demand connectivity verification.

RDI

Remote Defect Indication (RDI) is triggered by failure of proactive CC-CV, which is BFD based. For fast RDI initiation, RDI SHOULD be initiated and handled by hardware if BFD is handled in forwarding hardware. [RFC6428] provides an extension for BFD that includes the RDI indication in the BFD format and a specification of how this indication is to be used.

Route Tracing

[RFC6426] specifies that the LSP ping enhancements for MPLS-TP on-demand connectivity verification include information on the use of LSP ping for route tracing of an MPLS-TP path.

Alarm Reporting

[RFC6427] describes the details of a new protocol supporting Alarm Indication Signal, Link Down Indication, and fault management. This functionality SHOULD be supported in forwarding hardware on high speed interfaces.

Lock Instruct

Lock instruct is initiated on-demand and therefore need not be implemented in forwarding hardware. [RFC6435] defines a lock instruct protocol.

Lock Reporting

[RFC6427] covers lock reporting. Lock reporting need not be implemented in forwarding hardware.

Diagnostic

[RFC6435] defines protocol support for loopback. Loopback initiation is on-demand and therefore need not be implemented in forwarding hardware. Loopback of packet traffic SHOULD be implemented in forwarding hardware on high speed interfaces.

Packet Loss and Delay Measurement

[RFC6374] and [RFC6375] define a protocol and profile for packet loss measurement (LM) and delay measurement (DM). LM requires a very accurate capture and insertion of packet and byte counters when a packet is transmitted and capture of packet and byte counters when a packet is received. This capture and insertion MUST be implemented in forwarding hardware for LM OAM to be sufficiently accurate. DM requires very accurate capture and insertion of a timestamp on transmission and capture of timestamp when a packet is received. This timestamp capture and insertion MUST be implemented in forwarding hardware for DM OAM to be sufficiently accurate.

See Section 2.6.2 for discussion of hardware support necessary for BFD and LSP Ping.

CC-CV and alarm reporting is tied to protection and therefore SHOULD be supported in forwarding hardware in order to provide protection for a large number of affected LSP within target response intervals. Since CC-CV is supported by BFD, for MPLS-TP, BFD SHOULD be supported in forwarding hardware.

2.6.5. MPLS OAM and Layer-2 OAM Interworking

[RFC6670] provides the reasons for selecting a single MPLS-TP OAM solution and examines the consequences were ITU-T to develop a second OAM solution that is based on Ethernet encodings and mechanisms.

[RFC6310] and [I-D.ietf-pwe3-mpls-eth-oam-iwk] specifies the mapping of defect states between many types of hardware Attachment Circuits (ACs) and associated Pseudowires (PWs). This functionality SHOULD be supported in forwarding hardware.

An MPLS OAM implementation SHOULD interwork with the underlying server layer and provide a means to interwork with a client layer. Where MPLS hierarchy is used both the client and server layer may be MPLS or MPLS-TP. Where the server layer is a Layer-2, such as Ethernet, PPP over SONET/SDH, or GFP over OTN, interwork among layers is also required. For high speed interfaces, this interworking SHOULD be supported in forwarding hardware.

2.6.6. Extent of OAM Support by Hardware

Some OAM functionality must be supported in forwarding hardware while other OAM functionality must be entirely implemented in forwarding hardware.

Where possible, implementation in forwarding hardware should be in

programmable hardware such that if standards are later changed or extended these changes are likely to be accommodated with hardware reprogramming rather than replacement.

Some functions must be implemented in dedicated forwarding hardware. Examples include packet and byte counters needed for LM OAM as well as needed for management protocols. Similarly the capture and insertion of packet and byte counts or timestamps needed for transmitted LM or DM or time synchronization packets MUST be implemented in forwarding hardware to support accurate OAM.

Some functions must be supported in forwarding hardware but may make use of an external general purpose processor if performance criteria can be met. For example origination of AIS to client layers may be triggered by CC-CV server layer hardware but expansion to a large number of client LSP may occur in a general purpose processor. Some forwarding hardware supports one or more on-chip general purpose processors which may be well suited for such a role.

The customer (system supplier or provider) should not dictate design, but should independently validate target functionality and performance. However, it is not uncommon for service providers and system implementors to insist on reviewing design details (under NDA) due to past experiences with suppliers and to reject suppliers who are unwilling to provide details.

2.7. Number and Size of Flows

Service provider networks may carry up to hundreds of millions of flows on 10 Gb/s links. Most flows are very short lived, many under a second. A subset of the flows are low capacity and somewhat long lived. When Internet traffic dominates capacity a very small subset of flows are high capacity and/or very long lived.

Two types of limitations with regard to number and size of flows have been observed.

1. Some hardware cannot handle some very large flows because of internal paths which are limited, such as per packet backplane paths or paths internal or external to chips such as buffer memory paths. Such designs can handle aggregates of smaller flows. Some hardware with acknowledged limitations has been successfully deployed but may be increasingly problematic if the capacity of large microflows in deployed networks continues to grow.
2. Some hardware approaches cannot handle a large number of flows, or a large number of large flows due to attempting to count per

flow, rather than deal with aggregates of flows. Hash techniques scale with regard to number of flows due to a fixed hash size with many flows falling into the same hash bucket. Techniques that identify individual flows have been implemented but have never successfully deployed for Internet traffic.

3. Questions for Suppliers

The following questions should be asked of a supplier. These questions are grouped into broad categories. The questions themselves are intended to be an open ended question to the supplier. The tests in Section 4 are intended to verify whether the supplier disclosed any compliance or performance limitations completely and accurately.

Basic Compliance

- Q#1 Can the implementation forward packets with an arbitrarily large stack depth? What limitations exist, and under what circumstances do further limitations come into play (such as high packet rate or specific features enabled or specific types of packet processing)? See Section 2.1.
- Q#2 Is the entire set of basic MPLS functionality described in Section 2.1 supported?
- Q#3 Are the set of MPLS reserved labels handled correctly and with adequate performance? See Section 2.1.1.
- Q#4 Are mappings of label value and TC to PHB handled correctly, including RFC3270 L-LSP mappings and RFC4124 CT mappings to PHB? See Section 2.1.2.
- Q#5 Is time synchronization adequately supported in forwarding hardware?
 - a. Are both PTP and NTP formats supported?
 - b. Is the accuracy of timestamp insertion and incoming stamping sufficient?See Section 2.1.3.
- Q#6 Is link bundling supported?
 - a. Can LSP be pinned to specific components?

- b. Is the "all-ones" component link supported?

See Section 2.1.5.

Q#7 Is MPLS hierarchy supported?

- a. Are both PHP and UHP supported? What limitations exist on the number of POP operations with UHP?
- b. Are the pipe, short-pipe, and uniform models supported? Are TTL and TC values updated correctly at egress where applicable?

See Section 2.1.6

Q#8 Are pseudowire sequence numbers handled correctly? See Section 2.1.8.1.

Q#9 Is VPN LER functionality handled correctly and without performance issues? See Section 2.1.9.

Q#10 Is MPLS multicast (P2MP and MP2MP) handled correctly?

- a. Are packets dropped on uncongested outputs if some outputs are congested?
- b. Is performance limited in high fanout situations?

See Section 2.2.

Basic Performance

Q#11 Can very small packets be forwarded at full line rate on all interfaces indefinitely? What limitations exist, and under what circumstances do further limitations come into play (such as specific features enabled or specific types of packet processing)?

Q#12 Customers must decide whether to relax the prior requirement and to what extent. If the answer to the prior question indicates that limitations exist, then:

- a. What is the smallest packet size where full line rate forwarding can be supported?
- b. What is the longest burst of full rate small packets that can be supported?

Specify circumstances (such as specific features enabled or specific types of packet processing) often impact these rates and burst sizes.

- Q#13 How many POP operations can be supported along with a SWAP operation at full line rate while maintaining per LSP packet and byte counts for each POP and SWAP? This requirement is particularly relevant for MPLS-TP.
- Q#14 How many PUSH labels can be supported. While this limitation is rarely an issue, it applies to both PHP and UHP, unlike the POP limit which applies to UHP.
- Q#15 For a worst case where all packets arrive on one LSP, what is the counter overflow time? Are any means provided to avoid polling all counters at short intervals? This applies to both MPLS and MPLS-TP.

Multipath Capabilities and Performance

Multipath capabilities and performance do not apply to MPLS-TP but apply to MPLS and apply if MPLS-TP is carried in MPLS.

- Q#16 How are large microflows accommodated? Is there active management of the hash space mapping to output ports? See Section 2.4.2.
- Q#17 How many MPLS labels can be included in a hash based on the MPLS label stack?
- Q#18 Is packet rate performance decreased beyond some number of labels?
- Q#19 Can the IP header and payload information below the MPLS stack be used in the hash? If so, which IP fields, payload types and payload fields are supported?
- Q#20 At what maximum MPLS label stack depth can Bottom of Stack and an IP header appear without impacting packet rate performance?
- Q#21 Are reserved labels excluded from the label stack hash? They MUST be excluded.
- Q#22 How is multipath performance affected by very large flows or an extremely large number of flows, or by very short lived flows? See Section 2.7.

Pseudowire Capabilities and Performance

- Q#23 Is the pseudowire control word supported?
- Q#24 What is the maximum rate of pseudowire encapsulation and decapsulation? Apply the same questions as in Base Performance for any packet based pseudowire such as IP VPN or Ethernet.
- Q#25 Does inclusion of a pseudowire control word impact performance?
- Q#26 Are flow labels supported?
- Q#27 If so, what fields are hashed on for the flow label for different types of pseudowires?
- Q#28 Does inclusion of a flow label impact performance?

Entropy Label Support and Performance

- Q#29 Can an entropy label be added when acting as in ingress LER and can it be removed when acting as an egress LER?
- Q#30 If so, what fields are hashed on for the entropy label?
- Q#31 Does adding or removing an entropy label impact packet rate performance?
- Q#32 Can an entropy label be detected in the label stack, used in the hash, and properly terminate the search for further information to hash on?
- Q#33 Does using an entropy label have any negative impact on performance? It should have no impact or a positive impact.

OAM and DoS Protection

- Q#34 For each control and management plane protocol in use, what measures are taken to provide DoS attack hardenning? Have DoS attack tests been performed? Can compromise of an internal computer on a management subnet be leveraged for any form of attack including DoS attack?

Q#35 What OAM proactive and on-demand mechanisms are supported? What performance limits exist under high proactive monitoring rates? Can excessively high proactive monitoring rates impact control plane performance or cause control plane instability? Ask these questions for each of the following.

- a. MPLS OAM
- b. Pseudowire OAM
- c. MPLS-TP OAM
- d. Layer-2 OAM Interworking

See Section 2.6.

4. Forwarding Compliance and Performance Testing

Packet rate performance of equipment supporting a large number of 10 Gb/s or 100 Gb/s links is not possible using desktop computers or workstations. The use of high end workstations as a source of test traffic was barely viable 20 years ago, but is no longer at all viable. Though custom microcode has been used on specialized router forwarding cards to serve the purpose of generating test traffic and measuring it, for the most part performance testing will require specialized test equipment. There are multiple sources of suitable equipment.

The set of tests listed here do not correspond one-to-one to the set of questions in Section 3. The same categorization is used and these tests largely serve to validate answers provided to the prior questions, and can also provide answers where a supplier is unwilling to disclose compliance or performance.

Performance testing is the domain of the IETF Benchmark Methodology Working Group (BMWG). Below are brief descriptions of conformance and performance tests. Some very basic tests are specified in [RFC5695] which partially cover only the basic performance test T#3.

The following tests should be performed by the systems designer, or deployer, or performed by the supplier on their behalf if it is not practical for the potential customer to perform the tests directly. These tests are grouped into broad categories.

Basic Compliance

- T#1 Test forwarding at a high rate for packets with varying number of label entries. While packets with more than a dozen label entries are unlikely to be used in any practical scenario today, it is useful to know if limitations exists.
- T#2 For each of the questions listed under "Basic Compliance" in Section 3, verify the claimed compliance. For any functionality considered critical to a deployment, where applicable performance using each capability under load should be verified in addition to basic compliance.

Basic Performance

- T#3 Test packet forwarding at full line rate with small packets. See [RFC5695]. The most likely case to fail is the smallest packet size. Also test with packet sizes in four byte increments ranging from payload sizes of 40 to 128 bytes.
- T#4 If the prior tests did not succeed for all packet sizes, then perform the following tests.
 - a. Increase the packet size by 4 bytes until a size is found that can be forwarded at full rate.
 - b. Inject bursts of consecutive small packets into a stream of larger packets. Allow some time for recovery between bursts. Increase the number of packets in the burst until packets are dropped.
- T#5 Send test traffic where a SWAP operation is required. Also set up multiple LSP carried over other LSP where the device under test (DUT) is the egress of these LSP. Create test packets such that the SWAP operation is performed after POP operations, increasing the number of POP operations until forwarding of small packets at full line rate can no longer be supported. Also check to see how many POP operations can be supported before the full set of counters can no longer be maintained. This requirement is particularly relevant for MPLS-TP.
- T#6 Send all traffic on one LSP and see if the counters become inaccurate. Often counters on silicon are much smaller than the 64 bit packet and byte counters in IETF MIB. System developers should consider what counter polling rate is necessary to maintain accurate counters and whether those polling rates are practical. Relevant MIBs for MPLS are

discussed in [RFC4221] and [RFC6639].

Multipath Capabilities and Performance

Multipath capabilities do not apply to MPLS-TP but apply to MPLS and apply if MPLS-TP is carried in MPLS.

- T#7 Send traffic at a rate well exceeding the capacity of a single multipath component link, and where entropy exists only below the top of stack. If only the top label is used this test will fail immediately.
- T#8 Move the labels with entropy down in the stack until either the full forwarding rate can no longer be supported or most or all packets try to use the same component link.
- T#9 Repeat the two tests above with the entropy contained in IP headers or IP payload fields below the label stack rather than in the label stack. Test with the set of IP headers or IP payload fields considered relevant to the deployment or to the target market.
- T#10 Determine whether traffic that contains a pseudowire control word is interpreted as IP traffic. Information in the payload MUST NOT be used in the load balancing if the first nibble of the packet is not 4 or 6 (IPv4 or IPv6).
- T#11 Determine whether reserved labels are excluded from the label stack hash. They MUST be excluded.
- T#12 Perform testing in the presence of combinations of:
 - a. Very large microflows.
 - b. Relatively short lived high capacity flows.
 - c. Extremely large numbers of flows.
 - d. Very short lived small flows.

Pseudowire Capabilities and Performance

- T#13 Ensure that pseudowire can be set up with a pseudowire label and pseudowire control word added at ingress and the pseudowire label and pseudowire control word removed at egress.

- T#14 For pseudowire that contains variable length payload packets, repeat performance tests listed under "Basic Performance" for pseudowire ingress and egress functions.
- T#15 Repeat pseudowire performance tests with and without a pseudowire control word.
- T#16 Determine whether pseudowire can be set up with a pseudowire label, flow label, and pseudowire control word added at ingress and the pseudowire label, flow label, and pseudowire control word removed at egress.
- T#17 Determine which payload fields are used to create the flow label and whether the set of fields and algorithm provide sufficient entropy for load balancing.
- T#18 Repeat pseudowire performance tests with flow labels included.

Entropy Label Support and Performance

- T#19 Determine whether entropy labels can be added at ingress and removed at egress.
- T#20 Determine which fields are used to create an entropy label. Labels further down in the stack, including entropy labels further down and IP headers or IP payload fields where applicable should be used. Determine whether the set of fields and algorithm provide sufficient entropy for load balancing.
- T#21 Repeat performance tests under "Basic Performance" when entropy labels are used, where ingress or egress is the device under test (DUT).
- T#22 Determine whether an ELI is detected when acting as a midpoint LSR and whether the search for further information on which to base the load balancing is used. Information below the entropy label SHOULD NOT be used.
- T#23 Ensure that the Entropy Label Indicator and Entropy Label (ELI and EI) are removed from the label stack during UHP and PHP operations.
- T#24 Insure that operations on the TC field when adding and removing Entropy Label are correctly carried out. If TC is changed during a SWAP operation, the ability to transfer that change MUST be provided. The ability to suppress the

transfer of TC MUST also be provided. See "pipe", "short pipe", and "uniform" models in [RFC3443].

- T#25 Repeat performance tests for midpoint LSR with entropy labels found at various label stack depths.

DoS Protection

- T#26 Actively attack LSR under high protocol churn load and determine control plane performance impact or successful DoS under test conditions. Specifically test for the following.
- a. TCP SYN attack against control plane and management plane protocols using TCP, including CLI access (typically SSH protected login), NETCONF, etc.
 - b. High traffic volume attack against control plane and management plane protocols not using TCP.
 - c. Attacks which can be performed from a compromised management subnet computer, but not one with authentication keys.
 - d. Attacks which can be performed from a compromised peer within the control plane (internal domain and external domain). Assume that per peering keys and per router ID keys rather than network wide keys are in use.

See Section 2.6.1.

OAM Capabilities and Performance

- T#27 Determine maximum sustainable rates of BFD traffic. If BFD requires CPU intervention, determine both maximum rates and CPU loading when multiple interfaces are active.
- T#28 Verify LSP Ping and LSP Traceroute capability.
- T#29 Determine maximum rates of MPLS-TP CC-CV traffic. If CC-CV requires CPU intervention, determine both maximum rates and CPU loading when multiple interfaces are active.
- T#30 Determine MPLS-TP DM precision.

T#31 Determine MPLS-TP LM accuracy.

T#32 Verify MPLS-TP AIS/RDI and PSC functionality, protection speed, and AIS/RDI notification speed when a large number of Management Entities (ME) must be notified with AIS/RDI.

The tests in the "Basic Performance" section of the above list should be repeated under various conditions to retest basic performance when critical capabilities are enabled. Complete repetition of the performance tests enabling each capability and combinations of capabilities would be very time intensive, therefore a reduced set of performance tests can be used to gauge the impact of enabling specific capabilities.

5. Acknowledgements

Numerous very useful comments have been received in private email. Some of these contributions are acknowledged here, approximately in chronologic order.

Paul Doolan provided a brief review resulting in a number of clarifications, most notably regarding on-chip vs. system buffering, 100 Gb/s link speed assumptions in the 150 Mpps figure, and handling of large microflows. Pablo Frank reminded us of the sawtooth effect in PPS vs. packet size graphs, prompting the addition of a few paragraphs on this. Comments from Lou Berger at IETF-85 prompted the addition of Section 2.7.

Valuable comments were received on the BMWG mailing list. Jay Karthik pointed out extraneous methodology hints that belong in an appendix or should be removed.

Nabil Bitar pointed out the need to cover QoS (Differentiated Services), MPLS multicast (P2MP and MP2MP), and MPLS-TP OAM. Nabil also provided a number of clarifications to the questions and tests in Section 3 and Section 4.

6. IANA Considerations

This memo includes no request to IANA.

7. Security Considerations

This document reviews forwarding behavior specified elsewhere and points out compliance and performance requirements. As such it

introduces no new security requirements or concerns.

Knowledge of potential performance shortcomings may serve to help new implementations avoid pitfalls. It is unlikely that such knowledge could be the basis of new denial of service as these pitfalls are already widely known in the service provider community and among leading equipment suppliers. In practice extreme data and packet rate are needed to affect existing equipment and networks that may be still vulnerable due to failure to implement adequate protection and make this type of denial of service unlikely and make undetectable denial of service of this type impossible.

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.
- [RFC3032] Rosen, E., Tappan, D., Fedorkow, G., Rekhter, Y., Farinacci, D., Li, T., and A. Conta, "MPLS Label Stack Encoding", RFC 3032, 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.
- [RFC3270] Le Faucheur, F., Wu, L., Davie, B., Davari, S., Vaananen, P., Krishnan, R., Cheval, P., and J. Heinanen, "Multi-Protocol Label Switching (MPLS) Support of Differentiated Services", RFC 3270, May 2002.
- [RFC3443] Agarwal, P. and B. Akyol, "Time To Live (TTL) Processing in Multi-Protocol Label Switching (MPLS) Networks", RFC 3443, January 2003.
- [RFC4090] Pan, P., Swallow, G., and A. Atlas, "Fast Reroute Extensions to RSVP-TE for LSP Tunnels", RFC 4090, May 2005.
- [RFC4182] Rosen, E., "Removing a Restriction on the use of MPLS Explicit NULL", RFC 4182, September 2005.
- [RFC4201] Kompella, K., Rekhter, Y., and L. Berger, "Link Bundling in MPLS Traffic Engineering (TE)", RFC 4201, October 2005.
- [RFC4385] Bryant, S., Swallow, G., Martini, L., and D. McPherson,

"Pseudowire Emulation Edge-to-Edge (PWE3) Control Word for Use over an MPLS PSN", RFC 4385, February 2006.

- [RFC5129] Davie, B., Briscoe, B., and J. Tay, "Explicit Congestion Marking in MPLS", RFC 5129, January 2008.
- [RFC5586] Bocci, M., Vigoureux, M., and S. Bryant, "MPLS Generic Associated Channel", RFC 5586, June 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", RFC 6391, November 2011.
- [RFC6790] Kompella, K., Drake, J., Amante, S., Henderickx, W., and L. Yong, "The Use of Entropy Labels in MPLS Forwarding", RFC 6790, November 2012.

8.2. Informative References

- [ACK-compression]
"Observations and Dynamics of a Congestion Control Algorithm: The Effects of Two-Way Traffic", Proc. ACM SIGCOMM, ACM Computer Communications Review (CCR) Vol 21, No 4, 1991, pp.133-147., 1991.
- [ATM-EPD-and-PPD]
"Dynamics of TCP Traffic over ATM Networks", IEEE Journal of Special Areas of Communication Vol 13 No 4, May 1995, pp. 633-641., May 1995.
- [I-D.ietf-pwe3-mpls-eth-oam-iwk]
Mohan, D., Bitar, N., and A. Sajassi, "MPLS and Ethernet OAM Interworking", draft-ietf-pwe3-mpls-eth-oam-iwk-07 (work in progress), January 2013.
- [I-D.ietf-tictoc-1588overmpls]
Davari, S., Oren, A., Bhatia, M., Roberts, P., and L. Montini, "Transporting Timing messages over MPLS Networks", draft-ietf-tictoc-1588overmpls-03 (work in progress), October 2012.
- [RFC0791] Postel, J., "Internet Protocol", STD 5, RFC 791, September 1981.
- [RFC2474] Nichols, K., Blake, S., Baker, F., and D. Black, "Definition of the Differentiated Services Field (DS Field) in the IPv4 and IPv6 Headers", RFC 2474,

December 1998.

- [RFC2475] Blake, S., Black, D., Carlson, M., Davies, E., Wang, Z., and W. Weiss, "An Architecture for Differentiated Services", RFC 2475, December 1998.
- [RFC2597] Heinanen, J., Baker, F., Weiss, W., and J. Wroclawski, "Assured Forwarding PHB Group", RFC 2597, June 1999.
- [RFC3031] Rosen, E., Viswanathan, A., and R. Callon, "Multiprotocol Label Switching Architecture", RFC 3031, January 2001.
- [RFC3429] Ohta, H., "Assignment of the 'OAM Alert Label' for Multiprotocol Label Switching Architecture (MPLS) Operation and Maintenance (OAM) Functions", RFC 3429, November 2002.
- [RFC3985] Bryant, S. and P. Pate, "Pseudo Wire Emulation Edge-to-Edge (PWE3) Architecture", RFC 3985, March 2005.
- [RFC4110] Callon, R. and M. Suzuki, "A Framework for Layer 3 Provider-Provisioned Virtual Private Networks (PPVPNs)", RFC 4110, July 2005.
- [RFC4124] Le Faucheur, F., "Protocol Extensions for Support of Diffserv-aware MPLS Traffic Engineering", RFC 4124, June 2005.
- [RFC4206] Kompella, K. and Y. Rekhter, "Label Switched Paths (LSP) Hierarchy with Generalized Multi-Protocol Label Switching (GMPLS) Traffic Engineering (TE)", RFC 4206, October 2005.
- [RFC4221] Nadeau, T., Srinivasan, C., and A. Farrel, "Multiprotocol Label Switching (MPLS) Management Overview", RFC 4221, November 2005.
- [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.
- [RFC4379] Kompella, K. and G. Swallow, "Detecting Multi-Protocol Label Switched (MPLS) Data Plane Failures", RFC 4379, February 2006.
- [RFC4664] Andersson, L. and E. Rosen, "Framework for Layer 2 Virtual Private Networks (L2VPNs)", RFC 4664, September 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.
- [RFC4928] Swallow, G., Bryant, S., and L. Andersson, "Avoiding Equal Cost Multipath Treatment in MPLS Networks", BCP 128, RFC 4928, June 2007.
- [RFC4950] Bonica, R., Gan, D., Tappan, D., and C. Pignataro, "ICMP Extensions for Multiprotocol Label Switching", RFC 4950, August 2007.
- [RFC5082] Gill, V., Heasley, J., Meyer, D., Savola, P., and C. Pignataro, "The Generalized TTL Security Mechanism (GTSM)", RFC 5082, October 2007.
- [RFC5085] Nadeau, T. and C. Pignataro, "Pseudowire Virtual Circuit Connectivity Verification (VCCV): A Control Channel for Pseudowires", RFC 5085, December 2007.
- [RFC5332] Eckert, T., Rosen, E., Aggarwal, R., and Y. Rekhter, "MPLS Multicast Encapsulations", RFC 5332, August 2008.
- [RFC5462] Andersson, L. and R. Asati, "Multiprotocol Label Switching (MPLS) Label Stack Entry: "EXP" Field Renamed to "Traffic Class" Field", RFC 5462, February 2009.
- [RFC5695] Akhter, A., Asati, R., and C. Pignataro, "MPLS Forwarding Benchmarking Methodology for IP Flows", RFC 5695, November 2009.
- [RFC5860] Vigoureux, M., Ward, D., and M. Betts, "Requirements for Operations, Administration, and Maintenance (OAM) in MPLS Transport Networks", RFC 5860, May 2010.
- [RFC5880] Katz, D. and D. Ward, "Bidirectional Forwarding Detection (BFD)", RFC 5880, June 2010.
- [RFC5884] Aggarwal, R., Kompella, K., Nadeau, T., and G. Swallow, "Bidirectional Forwarding Detection (BFD) for MPLS Label Switched Paths (LSPs)", RFC 5884, June 2010.
- [RFC5885] Nadeau, T. and C. Pignataro, "Bidirectional Forwarding Detection (BFD) for the Pseudowire Virtual Circuit Connectivity Verification (VCCV)", RFC 5885, June 2010.
- [RFC5905] Mills, D., Martin, J., Burbank, J., and W. Kasch, "Network

Time Protocol Version 4: Protocol and Algorithms Specification", RFC 5905, June 2010.

- [RFC6310] Aissaoui, M., Busschbach, P., Martini, L., Morrow, M., Nadeau, T., and Y(J). Stein, "Pseudowire (PW) Operations, Administration, and Maintenance (OAM) Message Mapping", RFC 6310, July 2011.
- [RFC6371] Busi, I. and D. Allan, "Operations, Administration, and Maintenance Framework for MPLS-Based Transport Networks", RFC 6371, September 2011.
- [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.
- [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.
- [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.
- [RFC6426] Gray, E., Bahadur, N., Boutros, S., and R. Aggarwal, "MPLS On-Demand Connectivity Verification and Route Tracing", RFC 6426, November 2011.
- [RFC6427] Swallow, G., Fulignoli, A., Vigoureux, M., Boutros, S., and D. Ward, "MPLS Fault Management Operations, Administration, and Maintenance (OAM)", RFC 6427, November 2011.
- [RFC6428] Allan, D., Swallow Ed. , G., and J. Drake Ed. , "Proactive Connectivity Verification, Continuity Check, and Remote Defect Indication for the MPLS Transport Profile", RFC 6428, November 2011.
- [RFC6435] Boutros, S., Sivabalan, S., Aggarwal, R., Vigoureux, M.,

and X. Dai, "MPLS Transport Profile Lock Instruct and Loopback Functions", RFC 6435, November 2011.

- [RFC6438] Carpenter, B. and S. Amante, "Using the IPv6 Flow Label for Equal Cost Multipath Routing and Link Aggregation in Tunnels", RFC 6438, November 2011.
- [RFC6478] Martini, L., Swallow, G., Heron, G., and M. Bocci, "Pseudowire Status for Static Pseudowires", RFC 6478, May 2012.
- [RFC6639] King, D. and M. Venkatesan, "Multiprotocol Label Switching Transport Profile (MPLS-TP) MIB-Based Management Overview", RFC 6639, June 2012.
- [RFC6669] Sprecher, N. and L. Fang, "An Overview of the Operations, Administration, and Maintenance (OAM) Toolset for MPLS-Based Transport Networks", RFC 6669, July 2012.
- [RFC6670] Sprecher, N. and KY. Hong, "The Reasons for Selecting a Single Solution for MPLS Transport Profile (MPLS-TP) Operations, Administration, and Maintenance (OAM)", RFC 6670, July 2012.
- [RFC6720] Pignataro, C. and R. Asati, "The Generalized TTL Security Mechanism (GTSM) for the Label Distribution Protocol (LDP)", RFC 6720, August 2012.
- [RFC6829] Chen, M., Pan, P., Pignataro, C., and R. Asati, "Label Switched Path (LSP) Ping for Pseudowire Forwarding Equivalence Classes (FECs) Advertised over IPv6", RFC 6829, January 2013.

Appendix A. Organization of References Section

The References section is split into Normative and Informative subsections. References that directly specify forwarding encapsulations or behaviors are listed as normative. References which describe signaling only, though normative with respect to signaling, are listed as informative. They are informative with respect to MPLS forwarding.

Authors' Addresses

Curtis Villamizar (editor)
Outer Cape Cod Network Consulting, LLC

Email: curtis@occnc.com

Kireeti Kompella
Contrail Systems

Email: kireeti.kompella@gmail.com

Shane Amante
Level 3 Communications, Inc.
1025 Eldorado Blvd
Broomfield, CO 80021

Email: shane@level3.net

Andrew Malis
Verizon
60 Sylvan Road
Waltham, MA 02451

Phone: +1 781-466-2362
Email: andrew.g.malis@verizon.com

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

Email: cpignata@cisco.com

MPLS
Internet-Draft
Intended status: Standards Track
Expires: May 17, 2013

C. Villamizar, Ed.
Outer Cape Cod Network
Consulting
November 13, 2012

Multipath Extensions for MPLS Traffic Engineering
draft-villamizar-mpls-multipath-extn-00

Abstract

Extensions to OSPF-TE, ISIS-TE, and RSVP-TE are defined in support of carrying LSP with strict packet ordering requirements over multipath and carrying LSP with strict packet ordering requirements within LSP without violating requirements to maintain packet ordering. LSP with strict packet ordering requirements include MPLS-TP LSP.

OSPF-TE and ISIS-TE extensions defined here indicate node and link capability regarding support for ordered aggregates of traffic, multipath traffic distribution, and abilities to support multipath load distribution differently per LSP.

RSVP-TE extensions either identifies an LSP as requiring strict packet order, or identifies an LSP as carrying one or more LSP that requires strict packet order further down in the label stack, or identifies an LSP as having no restrictions on packet ordering except the restriction to avoid reordering microflows. In addition an extension indicates whether the first nibble of payload will reliably indicate whether payload is IPv4, IPv6, or other type of payload, most notably pseudowire using a pseudowire control word.

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 May 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 | 4 |
| 1.1. Architecture Summary | 4 |
| 1.2. Requirements Language | 5 |
| 1.3. Definitions | 5 |
| 2. Protocol Extensions | 6 |
| 2.1. Multipath Node Capability sub-TLV | 6 |
| 2.2. Multipath Link Capability TLV | 9 |
| 2.3. LSP Multipath Attributes TLV | 9 |
| 3. Protocol Mechanisms | 12 |
| 3.1. OSPF-TE and ISIS-TE Advertisement | 12 |
| 3.1.1. Node Capability Advertisement | 12 |
| 3.1.2. Link Capability Advertisement | 12 |
| 3.1.3. Setting Max Depth and IP Depth | 12 |
| 3.1.4. Advertising Multipath as Link Bundling | 13 |
| 3.1.5. Hierarchical LSP Link Advertisement | 13 |
| 3.1.6. Advertisement of Legacy Multipath | 14 |
| 3.2. RSVP-TE LSP Attributes | 15 |
| 3.2.1. LSP Contained Ordered Aggregates Flags | 15 |
| 3.2.2. LSP Attributes for Ordered Aggregates | 17 |
| 3.2.3. Attributes for LSP without Packet Ordering | 17 |
| 3.3. Path Computation Constraints | 20 |
| 3.3.1. Link Multipath Capabilities and Path Computation | 20 |
| 3.3.1.1. Path Computation with Ordering Constraints | 20 |
| 3.3.1.2. Path Computation with No Ordering Constraint | 21 |
| 3.3.1.3. Path Computation for MPLS containing MPLS-TP | 21 |
| 3.3.2. Link IP Capabilities and Path Computation | 21 |
| 3.3.2.1. LSP without Packet Ordering Requirements | 22 |
| 3.3.2.2. LSP with Ordering Requirements | 22 |
| 3.3.3. Link Depth Limitations and Path Computation | 23 |
| 4. Backwards Compatibility | 24 |
| 4.1. Legacy Multipath Behavior | 24 |
| 4.2. Networks without Multipath Extensions | 24 |
| 4.2.1. Networks with MP Capability on all Multipath | 24 |
| 4.2.2. Networks with OA Capability on all Multipath | 26 |
| 4.2.3. Legacy Networks with Mixed MP and OA Links | 26 |
| 4.3. Transition to Multipath Extension Support | 27 |
| 4.3.1. Simple Transitions | 27 |
| 4.3.2. More Challenging Transitions | 27 |
| 5. IANA Considerations | 28 |
| 6. Security Considerations | 28 |
| 7. References | 28 |
| 7.1. Normative References | 28 |
| 7.2. Informative References | 29 |
| Author's Address | 30 |

1. Introduction

Today the requirement to handle large aggregations of traffic, can be handled by a number of techniques which we will collectively call multipath. Multipath is very similar to composite link as defined in [ITU-T.G.800], except multipath specifically excludes inverse multiplexing. Some types of LSP, including but potentially not limited to MPLS-TP LSP, require strict packet ordering.

A means to support a MPLS-TP client layer over a MPLS server layer using MPLS Entropy Label is defined in [I-D.villamizar-mpls-multipath-use]. It is noted in [I-D.villamizar-mpls-multipath-use] that absent some protocol extensions, some limitations must be accepted.

This document defines protocol extensions which better supports using MPLS with multipath as a server layer for MPLS-TP, or to carry MPLS-TP directly over a network which makes use of multipath. Extensions are also applicable to MPLS when used in the presense of very large microflows or very large LSP which cannot be load split as a result of using MPLS Entropy Label [I-D.ietf-mpls-entropy-label].

1.1. Architecture Summary

Advertisements in a link state routing protocol, such as OSPF or ISIS, support a topology map known as a link state database (LSDB). When traffic engineering information is included in the LSDB the topology map is known as a TE-LSDB or traffic engineering database (TED).

A common MPLS LSP path computation is known as a constrained shortest path first computation (CSPF) (see [RFC3945]). Other algorithms may be used for path computation. Constraint-based routing was first introduced in [RFC2702]).

OSPF-TE or ISIS-TE extensions are defined in Section 2.1 and Section 2.2. OSPF-TE or ISIS-TE advertisements serve to populate the TE-LSDB and provide the basis for constraint-based routing path computation. Section 3.1 describes the use of OSPF-TE or ISIS-TE multipath extensions in routing advertisements.

RSVP-TE extensions are defined in Section 2.3. Section 3.2 describes the use of RSVP-TE extensions in setting up LSP including signaling constraints on LSP which contain other LSP which specify RSVP-TE extensions.

Section 3.3 describes the constraints on LSP path computation imposed by the advertised ordered aggregate and multipath capabilities of

links. Section 3.3.2 describes the constraints on LSP path computation imposed by link advertisements regarding use of IP headers in multipath traffic distribution. Section 3.3.3 describes the impact of label stack depth limitations.

1.2. Requirements Language

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

1.3. Definitions

Please refer to [I-D.villamizar-mpls-multipath-use].

Ordered Aggregate (OA)

An ordered aggregate (OA) requires that packets be delivered in the order in which they were received. Please refer to [RFC3260].

Microflow

A microflow is a single instance of an application-to-application flow. Please refer to [RFC2475]. Reordering packets within a microflow can cause service disruption. Please refer to [RFC2991].

Multipath Traffic Distribution

Multipath traffic distribution refers to the mechanism which distributes traffic among a set of component links or component lower layer paths which together comprise a multipath. No assumptions are made about the algorithms used in multipath traffic distribution. This document only discusses constraints of the type of information which can be used as the basis for multipath traffic distribution in specific circumstances.

The phrase "strict packet ordering requirements" refers to the requirement to deliver all packet in the order that they were received. The absence of strict packet ordering requirements does not imply total absence of packet ordering requirements. The requirement to avoid reordering traffic within any given microflow, as described in [RFC2991] applies to all traffic aggregates including all MPLS LSP.

The abbreviations ELI and EL are Entropy Label Indicator and Entropy Label, as defined in [I-D.ietf-mpls-entropy-label].

2. Protocol Extensions

This section defined protocol extensions to OSPF-TE, ISIS-TE, and RSVP-TE. Use of these extensions is described in Section 3 and Section 4.

Two capability sub-TLV are added to two TLV that are used in both OSPF-TE and ISIS-TE. The Multipath Node Capability sub-TLV is added to the Node Attribute TLV (see Section 2.1). The Multipath Link Capability TLV is added to the Interface_ID (see Section 2.2).

One TLV is added to the LSP_ATTRIBUTES object defined in [RFC5420].

2.1. Multipath Node Capability sub-TLV

The Node Attribute TLV is defined in [RFC5786]. A new sub-TLV, the Multipath Node Capability sub-TLV, is defined for inclusion in the Node Attribute TLV.

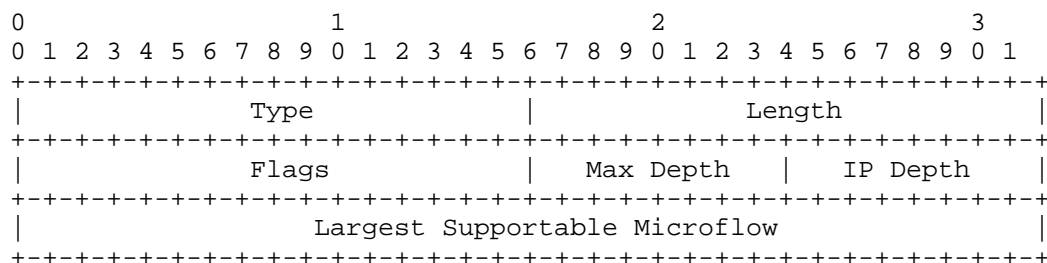


Figure 1: Multipath Capability Sub-TLV

The fields in the Multipath Capability sub-TLV are defined as follows.

Type

The Type field is assigned a value of IANA-TBD-1. The Type field is a two octet value.

Length

The Length field indicates the length of the sub-TLV in octets, excluding the Type and Length fields. The Length field is a two octet value.

Flags

The Flags field is a two octet (16 bit) value. The following single bit fields are assigned within this value, starting at the most significant bit, which is the bit transmitted first.

0x8000 Ordered Aggregate Enabled

Setting the Ordered Aggregate Enabled bit indicates that an LSP can be carried as an Ordered Aggregate Enabled on one or more links.

0x4000 Multipath Enabled

Setting the Multipath Enabled bit indicates that an LSP can be spread across component links at one or more multipath links.

0x2000 IPv4 Enabled Multipath

Setting the IPv4 Enabled Multipath bit indicates that the IPv4 header information can be used in multipath load balance. The Multipath Enabled bit must be set if the IPv4 Enabled Multipath bit is set.

0x1000 IPv6 Enabled Multipath

Setting the IP bit indicates that the IPv6 header information can be used in multipath load balance. The Multipath Enabled bit must be set if the IPv6 Enabled Multipath bit is set.

0x0800 UDP/IPv4 Multipath

Setting the UDP/IPv4 Multipath bit indicates that the UDP port numbers carried in UDP over IPv4 can be used in multipath load balance. The IPv4 Enabled Multipath bit must be set if UDP/IPv4 Multipath is set. If the IPv4 Enabled Multipath bit is set and the UDP/IPv4 Multipath bit is clear, then only source and destination IP addresses are used.

0x0400 UDP/IPv6 Multipath

Setting the UDP/IPv6 Multipath bit indicates that the UDP port numbers carried in UDP over IPv6 can be used in multipath load balance. The IPv6 Enabled Multipath bit must be set if UDP/IPv6 Multipath is set. The IPv6 Enabled Multipath bit must be set if UDP/IPv6 Multipath is set. If the IPv6 Enabled Multipath bit is set and the UDP/IPv6 Multipath bit is clear, then only source and destination IP addresses are used.

0x0200 TCP/IPv4 Multipath

Setting the TCP/IPv4 Multipath bit indicates that the TCP port numbers carried in TCP over IPv4 can be used in multipath load balance. The IPv4 Enabled Multipath bit must be set if TCP/IPv4 Multipath is set. If the IPv4 Enabled Multipath bit is set and the TCP/IPv4 Multipath bit is clear, then only source and destination IP addresses are used.

0x0100 TCP/IPv6 Multipath

Setting the TCP/IPv6 Multipath bit indicates that the TCP port numbers carried in TCP over IPv6 can be used in multipath load balance. The IPv6 Enabled Multipath bit must be set if TCP/IPv6 Multipath is set. The IPv6 Enabled Multipath bit must be set if TCP/IPv6 Multipath is set. If the IPv6 Enabled Multipath bit is set and the TCP/IPv6 Multipath bit is clear, then only source and destination IP addresses are used.

0x0080 Default to Multipath

Setting the Default to Multipath bit indicates that for an LSP which does not signal a desired behavior the traffic for that LSP will be spread across component links at one or more multipath links. If the Default to Multipath bit is not set, then an LSP which does not signal otherwise will be treated as an ordered aggregate.

0x0040 Default to IP/MPLS Multipath

Setting the Default to IP/MPLS Multipath indicates that for an LSP which does not signal a desired behavior, the IP header information will be used in the multipath load distribution. If the Default to IP/MPLS Multipath is clear it indicates that the the IP header information will not be used by default.

0x0020 Entropy Label Multipath

Setting the Entropy Label Multipath bit indicates that when multipath is enabled for a given LSP, if an Entropy Label Indicator (ELI) is found in the label stack, information below the Entropy Label (EL) will not be used in multipath load distribution.

0x0010 IP Optional Multipath

Setting the IP Optional Multipath bit indicates that when multipath is enabled for a given LSP, whether the IP header information is used in the multipath load distribution can be set on a per LSP basis.

The remaining bits in the Flags field are reserved.

Max Depth

The Max Depth field is a one octet field indicating the maximum label stack depth beyond which the multipath load distribution cannot make use of further label stack entries.

IP Depth

The IP Depth field is a one octet field indicating the maximum label stack depth beyond which the multipath load distribution cannot make use of IP information.

Largest Supportable Microflow

The Largest Supportable Microflow field is a IEEE 32 bit floating point value expressing in bytes/second. Any microflow which exceeds this capacity may experience either packet loss, or out-of-order delivery, or both.

The reserved bits in the Flags field MUST be set to zero and MUST be ignored unless implementing an extension which redefines one or more of the reserved bits. Any further extension which redefines one or more reserved Flags bit should maintain backwards compatibility with prior implementations.

2.2. Multipath Link Capability TLV

The Interface_ID is defined in [RFC3471]. The Interface_ID is updated in [RFC4201] to support a form of multipath known as Link Bundling.

A new TLV, the Multipath Link Capability TLV, is defined here. The Multipath Link Capability TLV is optionally included in the Interface_ID. The format of the Multipath Link Capability TLV is identical to the Multipath Node Capability sub-TLV defined in Section 2.1, with one exception. In the Multipath Link Capability TLV the Type field value is IANA-TBD-2.

If a Multipath Link Capability TLV is advertised for any link, then a Multipath Node Capability sub-TLV MUST be advertised for the node.

2.3. LSP Multipath Attributes TLV

The LSP_ATTRIBUTES object is defined in [RFC5420]. A new LSP Multipath Attributes TLV is defined for the LSP_ATTRIBUTES object. The TLV Type is IANA_TBD_3. The format is described below.

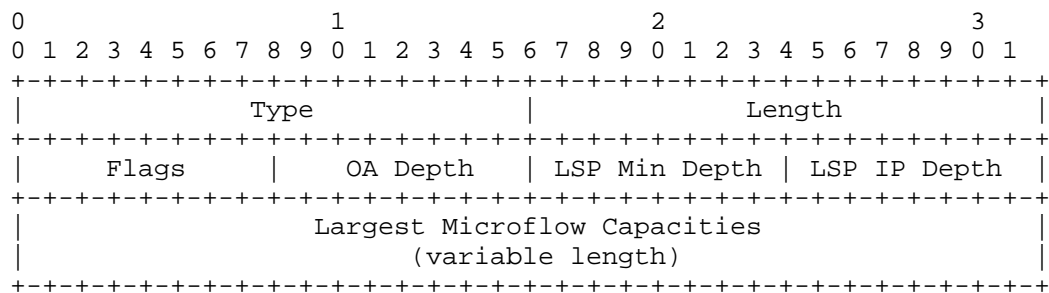


Figure 2: LSP Multipath Attributes TLV

The fields in the LSP Multipath Attributes TLV are defined as

follows.

Type

The Type field is assigned a value of IANA-TBD-3. The Type field is a two octet value.

Length

The Length field indicates the length of the sub-TLV in octets, excluding the Type and Length fields. The Length field is a two octet value.

Flags

The Flags field is a one octet (8 bit) value. The following single bit fields are assigned within this value, starting at the most significant bit, which is the bit transmitted first.

0x80 IP Multipath Allowed

Setting the IP Multipath Allowed bit indicates that it is safe to enable the use of a potential IP payload in the multipath traffic distribution.

0x40 May Contain IPv4

Setting the May Contain IPv4 bit indicates that IPv4 traffic may be contained within this LSP.

0x20 May Contain IPv6

Setting the May Contain IPv6 bit indicates that IPv6 traffic may be contained within this LSP.

0x02 Entropy Label Required

Setting the Entropy Label Used bit indicates that midpoint LSR MUST support ELI and EL in order to not violate packet ordering constraints of the LSP or of contained LSP.

0x01 Entropy Label Used

Setting the Entropy Label Used bit indicates that an ELI and EL is present in some or all label stack entries within this LSP.

The remaining bits in the Flags field are reserved.

OA Depth

The OA Depth field is set as follows

- 0 An OA Depth value of zero indicates that no ordered aggregates are carried within the LSP, except those which are protected from out of order delivery using Entropy Label.

- 1 An OA Depth value of one indicates that the LSP is an ordered aggregate of traffic (the LSP requires strict ordering of packets) and has protected packet ordering using ELI and EL.
- >1 An OA Depth value greater than one indicates that the LSP does not have strict packet ordering requirements but contains ordered aggregates at the label stack depth indicated or deeper and that the ordered aggregates are not protected using ELI and EL.

LSP Min Depth

The LSP Min Depth field indicates a minimal acceptable number of label used in multipath traffic distribution for the stated Largest Microflow Capacities field to be valid. If the LSP Min Depth field is set to zero this value is unknown. See Section 3.3.3.

LSP IP Depth

The LSP IP Depth field indicates a minimal label stack depth where using an IP header is necessary for the stated Largest Microflow Capacities field to be valid. If the LSP IP Depth field is set to zero this value is unknown. See Section 3.3.3.

Largest Microflow Capacities

The Largest Microflow Capacities field contains zero, one, two, or three IEEE 32 bit floating point values. Each value is a capacity expressed in bytes per second.

Largest LSE Microflow

The first value, the Largest LSE Microflow, is the capacity of the largest microflow if only the label stack entries are used in multipath traffic distribution. If a Largest LSE Microflow is not included, the LSP bandwidth request MUST be used.

Largest IP Microflow

The second value, the Largest IP Microflow, if present, is the capacity of the largest microflow if the label stack entries and any potential IP source and destination address are used in multipath traffic distribution. If the Largest IP Microflow is not included, the value of the Largest LSE Microflow MUST be used.

Largest L4 Microflow

The third, the Largest L4 Microflow, if present, is the capacity of the largest microflow if the label stack entries and any potential IP addresses and TCP or UDP port numbers are used in multipath traffic distribution. If a Largest L4 Microflow is not included, the value of the Largest IP

Microflow MUST be used.

3. Protocol Mechanisms

3.1. OSPF-TE and ISIS-TE Advertisement

Every compliant node MUST advertise exactly one Multipath Node Capability sub-TLV and MAY advertise zero or more Multipath Link Capability sub-TLV as needed.

Procedures for accommodating legacy forwarding capabilities and non-compliant nodes are discussed in Section 4.

3.1.1. Node Capability Advertisement

Every LSR which is adjacent to one or more multipath link MUST advertise a Multipath Node Capability sub-TLV (see Section 2.1). The capabilities advertised for the node SHOULD reflect the capabilities of the majority of multipath links adjacent to the node.

Every LSR which is not adjacent to any multipath links MUST advertise a Multipath Node Capability sub-TLV with both the Ordered Aggregate Enabled bit in Flags set and all other Flags bits clear. Both Max Depth and IP Depth can be set to zero. This advertisement identifies the LSR as one which is conformant but has no multipath links, allowing it to be distinguished from a non-conformant LSR with LAG or other multipath which may have to be avoided when determining a path for some LSP.

3.1.2. Link Capability Advertisement

For all of the links whose capability does not exactly match the Multipath Node Capability sub-TLV advertised by that same LSR, the LSR MUST advertise a Multipath Link Capability sub-TLV (see Section 2.2).

For all of the links whose capability does exactly match the Multipath Node Capability sub-TLV advertised by that same LSR, the LSR SHOULD NOT advertise a Multipath Link Capability sub-TLV (see Section 2.2). In this case the Multipath Link Capability TLV is redundant, but harmless.

3.1.3. Setting Max Depth and IP Depth

The Max Depth and IP Depth field are intended to capture architectural limits. Most forwarding hardware will only use a limited number of label entries in the multipath traffic

distribution. This limit is reflected in the Max Depth field. Most forwarding hardware will limit the number of label entries that it will look past before looking for an IP header to be used in the multipath traffic distribution. This limit is reflected in the IP Depth field.

3.1.4. Advertising Multipath as Link Bundling

All multipath links and FA for PSC LSP which traverse multipath links MAY be advertised as Link Bundles as defined in [RFC4201]. The settings of the Ordered Aggregate Enabled and Multipath Enabled fields indicate key capabilities of the multipath. Advertising the multipath as a link bundle can provide additional information, such as the capability to place LSP on individual components.

If the Multipath Enabled bit is set in the Multipath Link Capability TLV Flags, then the Maximum LSP Bandwidth in the Interface Identification TLV can be set in one of two ways.

1. If the desired behavior for legacy LSR acting as ingress is to limit LSP to the capacity of a single component link, then Maximum LSP Bandwidth SHOULD be set as specified in [RFC4201].
2. If the desired behavior for legacy LSR acting as ingress is to allow LSP to exceed the capacity of a single component link, then Maximum LSP Bandwidth MAY be set to a higher value, up to the value of Maximum Reservable Bandwidth. This would normally be done if the legacy LSR were known to be carrying traffic which is easily load split, such as typical Internet traffic.

LSR acting as ingress SHOULD ignore the Maximum LSP Bandwidth and MAY set up LSP with capacity up to the Maximum Reservable Bandwidth as long as microflows within the LSP will not exceed the Largest Supportable Microflow capacity.

3.1.5. Hierarchical LSP Link Advertisement

A PSC LSP, as defined in [RFC4206] and updated in [RFC6107], may carry other LSP. When signaling a PSC LSP expected characteristics of the contained traffic must be estimated. The FA advertised for the PSC LSP MUST reflect the broadest set of requirements the PSC LSP can carry. If the setup of an additional reservation would exceeded current capacity, a PSC LSP may be resigaled using make-before-break semantics ([RFC3209]).

For example, if it is expected that a PSC LSP will carry MPLS-TP LSP or other LSP with strict packet reordering requirements at some label depth, the minimum label stack depth at which an LSP with strict

packet reordering requirements can be carried must be signaled in the OA Depth field of the LSP Multipath Attributes TLV (see Section 2.3).

When the Forwarding Adjacency (FA) is advertised, the advertised Max Depth and IP Depth MUST be one less than the minimum of the Max Depth and IP Depth of any link that the PSC LSP traverses. The Max Depth and IP Depth are considered independently of each other. The reduction by one takes into account the PSC label. The Flags advertised for the FA MUST reflect the capabilities of the links along the path.

3.1.6. Advertisement of Legacy Multipath

An Ethernet LAG with no support for Entropy Label MUST have the Ordered Aggregate Enabled bit cleared and the Multipath Enabled bit set in the Multipath Link Capability TLV Flags. This indicates that a MPLS-TP compliant server layer cannot be supported and that LSP larger than the component links (LAG members) can be supported.

A Link Bundle that does not support the all-ones component defined in [RFC4201] MUST have the Ordered Aggregate Enabled bit set and the Multipath Enabled bit cleared in the Multipath Link Capability TLV Flags. This indicates that a MPLS-TP compliant server layer can be supported and that LSP larger than the component links cannot be supported.

A link bundle that can support either the pinning of a LSP to a single component link or the spreading of traffic across multiple component links MUST have the Ordered Aggregate Enabled bit set and the Multipath Enabled bit set in the Multipath Link Capability TLV Flags. This indicates that a MPLS-TP compliant server layer can be supported and that LSP larger than the component links can also be supported.

Any form of multipath that supports Entropy Label MUST have the Ordered Aggregate Enabled bit set and the Multipath Enabled bit set and the Entropy Label Multipath bit set in the Multipath Link Capability TLV Flags. Any form of multipath that does not support Entropy Label MUST have the Entropy Label Multipath bit cleared in the Multipath Link Capability TLV Flags.

The remaining bits in the Multipath Link Capability TLV Flags MUST be set according to the capability and configuration of the multipath or LSP.

3.2. RSVP-TE LSP Attributes

All LSR SHOULD advertise a LSP Multipath Attributes TLV with the RSVP-TE signaling for each LSP for which it is serving as ingress. If any legacy LSR remain in the network, it is easier to assign an acceptable default treatment for LSP signaled by those legacy LSR if the conforming LSR always send a LSP Multipath Attributes TLV.

There are two general cases, an LSP requires strict ordering of packets, or it doesn't. In the latter case the LSP may contain other LSP that require strict ordering and those must be protected from reordering using an Entropy Label as described in [I-D.villamizar-mpls-multipath-use]. These two cases are briefly described below.

Ordered Aggregates

LSP with strict packet order requirements MUST set the OA Depth field to one to indicate that the LSP MUST be treated as ordered aggregate. See Section 3.2.2.

LSP without Packet Ordering

LSP which do not have strict packet order requirements MUST only carry LSP whose requirements are reflected in the containing LSP Multipath Attributes TLV. See Section 3.2.3.

If an attempt is made to signal a contained LSP whose Ordered Aggregate Attributes TLV and LSP Multipath Attributes TLV cannot be supported by the containing (PSC) LSP, one of the two following actions MUST be taken.

1. The containing (PSC) LSP MAY be resigaled with appropriate TLVs to carry the new contained LSP using make-before-break semantics, then continue to forward the contained LSP PATH message if the containing (PSC) LSP is successfully updated.
2. The LSR MAY reject the contained LSP signaling by sending a PathErr message.

3.2.1. LSP Contained Ordered Aggregates Flags

The Flags field in the LSP Multipath Attributes TLV MUST be set as follows.

1. If the LSP may directly contain IPv4 traffic, then the May Contain IPv4 bit in the Flags field MUST be set.
2. If the LSP may directly contain IPv6 traffic, then the May Contain IPv6 bit in the Flags field MUST be set.

3. If the LSP contains an LSP which has the May Contain IPv4 bit in the Flags field then the May Contain IPv4 bit in the Flags field MUST be set in the containing LSP.
4. If the LSP contains an LSP which has the May Contain IPv6 bit in the Flags field then the May Contain IPv6 bit in the Flags field MUST be set in the containing LSP.
5. If the LSP may contain pseudowires that do not use a pseudowire control word [RFC4385], and may contain IPv4 or IPv6 traffic, then the IP Multipath Allowed bit in the Flags field MUST be cleared.
6. If the LSP is known to contain no pseudowires that do not use a pseudowire control word, then the IP Multipath Allowed bit in the Flags field SHOULD be set unless disallowed due to a contained LSP.
7. If an Entropy Label is added below the LSP label, then the Entropy Label Used bit MUST be set.
8. If the LSP contains any LSP with the IP Multipath Allowed bit in the Flags field clear, then the IP Multipath Allowed bit in the Flags field MUST be clear.

If the LSP does not contain other LSP, it may contain IP traffic and/or pseudowire that terminate on that LSR. If the LSP does not contain other LSP. The LER will know whether the LSP is used in an IP LER capacity. The LER will also know whether it terminates any pseudowire for a given LSP. The LER will also know if it is using Entropy Label for a given LSP and if it requires strict packet ordering for a given LSP. Therefore, when a LSP does not contain other LSP, then it is possible to accurately set the Flags field in the LSP Multipath Attributes TLV, as well the OA Depth, and LSP IP Depth fields.

If an LSP contains other LSP, and all of the contained include a LSP Multipath Attributes TLV, then it is still possible to accurately set the Flags field in the LSP Multipath Attributes TLV, as well the OA Depth, and LSP IP Depth fields. It is only when LSP contains other LSP that do not have a LSP Multipath Attributes TLV where default behavior has to be configured based on assumptions about LSP originated by the legacy LSR where there is a potential for those configured assumptions to be inaccurate.

See Section 4 for guidelines for handling LSP which contain LSP that do not have a LSP Multipath Attributes TLV. The most conservative approach in this case is to clear the IP Multipath Allowed bit and

set the May Contain IPv4 bit and the May Contain IPv6 bit, however this may not always be necessary.

3.2.2. LSP Attributes for Ordered Aggregates

An LSP with strict packet order requirements MUST always include a LSP Multipath Attributes TLV.

Most of the Flags in the LSP Multipath Attributes TLV can be set as described in Section 3.2.1. There are two cases which affect the setting of the remaining Flags bits.

Entropy Label not used

If an Entropy Label is not used, then the Entropy Label Used bit, the Entropy Label Required bit, and the IP Multipath Allowed bit MUST be cleared.

Entropy Label is used If an Entropy Label is used, then the Entropy Label Used bit, and the Entropy Label Required bit, and the IP Multipath Allowed bit MUST be set.

The OA Depth field MUST be set to one. The Min Depth MUST be set to one. The LSP IP Depth SHOULD be set to zero. The Largest Microflow Capacities field SHOULD be empty. The entire LSP is one microflow. The Largest Microflow Capacities field MAY contain one entry if there is some reason to do so, such as an LSP which may peak at capacity higher than the bandwidth reserved for the LSP. The Largest Microflow Capacities for an LSP SHOULD be configurable independently of the LSP bandwidth reservation.

3.2.3. Attributes for LSP without Packet Ordering

If an LSP does not have strict packet order constraints, then the LSR_ATTRIBUTE object SHOULD always include a LSP Multipath Attributes TLV.

Most of the Flags in the LSP Multipath Attributes TLV can be set as described in Section 3.2.1. There are two cases which affect the setting of the remaining Flags bits, the OA Depth field, the LSP Min Depth, and the LSP IP Depth field.

Entropy Label not used

- * The OA Depth MUST be either set to zero or set to a configured value that is greater than one, or set based on the contained LSP.

- * If the OA Depth is set to a configured value, then any setup attempt for a contained LSP with a depth greater than or equal to that value SHOULD be rejected and a PathErr message sent. Otherwise, if a setup attempt for a contained LSP with a depth greater than the current value included in the containing LSP OA Depth field, then the containing LSP MUST be rerouted with a OA Depth field value greater than any of the contained OA Depth field values.
- * The Entropy Label Used bit MUST be set if any contained LSP has the Entropy Label Used bit set.
- * The Entropy Label Required bit MUST be set if any contained LSP has the Entropy Label Required bit set.

Entropy Label is used

- * The OA Depth MUST be set to zero.
- * The Entropy Label Used bit MUST be set.
- * The Entropy Label Required bit MUST be set if any contained LSP has the Entropy Label Required bit set.
- * The Entropy Label Required bit MUST be set if any contained LSP has the OA Depth field set to a non-zero value.
- * The Entropy Label Required bit SHOULD be clear if there are no contained LSP has the OA Depth field set to a non-zero value and no contained LSP with the Entropy Label Required bit set. In this case the Entropy Label Required bit MAY be set by configuration, generally in anticipation of needing it in the future to carry other LSP.
- * LSP Min Depth field MUST be set to three if the Entropy Label Required bit is set.
- * If the Entropy Label Required bit is not set, then the LSP Min Depth field and LSP IP Depth field SHOULD be set to three if there are no contained LSP. The LSP Min Depth field and LSP IP Depth MAY be configured to a minimum value greater than three, generally in anticipation of needing it in the future to carry other LSP.
- * If the Entropy Label Required bit is not set, and there are contained LSP, then the LSP Min Depth field MUST be set to a value greater than three.

- * If the Entropy Label Required bit is not set, the LSP Min Depth field MUST be set to a value that is at least the sum of three plus the LSP Min Depth field in any contained LSP.
- * If the Entropy Label Required bit is not set, and either the May Contain IPv4 bit or the May Contain IPv6 bit is set, then the LSP IP Depth field MUST be set to a value of one or greater.
- * If the Entropy Label Required bit is not set, and any contained LSP has the May Contain IPv4 bit or the May Contain IPv6 bit is set, then the LSP IP Depth field MUST be set to a value of two or greater.
- * If the Entropy Label Required bit is not set, and any contained LSP has the LSP IP Depth field set to a value greater than one, then the LSP IP Depth field MUST be set to a value greater than the highest LSP IP Depth value set in any contained LSP.

The values of the LSP Min Depth field and the LSP IP Depth field MAY be constrained to upper limits by configuration. If an attempt to setup a contained LSP would result in exceeding one of these limits, then the LSR SHOULD reject the signaling attempt and send a PathErr message.

If Entropy Label is not used on the signaled LSP and there are no contained LSP, then the Largest LSE Microflow in the Largest Microflow Capacities field MUST be set to the requested bandwidth of the LSP. The optional Largest IP Microflow and Largest L4 Microflow SHOULD be included and MAY be set to configured minimum values.

If Entropy Label is not used on the signaled LSP an LSP that does not have strict packet order constraints contains other LSP, then the LSP Multipath Attributes TLV advertised by the set of contained LSP MUST be used to set the LSP Multipath Attributes TLV Largest Microflow Capacities values for LSP Multipath Attributes TLV. The value of Largest LSE Microflow, Largest IP Microflow, and Largest L4 Microflow in the LSP Multipath Attributes TLV of the containing LSP cannot be less than the maximum effective value of the same parameter for any contained LSP Multipath Attributes TLV.

If Entropy Label is used on the signaled LSP then the Largest LSE Microflow field is set according to the largest microflow that can result from computing the Entropy Label. This value is the greatest of a set of sources of entropy. A configured value MAY be used for IP, or it MAY be assumed that the largest interface carrying IP could carry a single microflow. For contained LSP which have the Entropy Label Used bit clear, the value in the Largest IP Microflow can be

used if the IP Multipath Allowed bit is set for that LSP and the LSR can make use of the IP information in the label stack. For contained LSP which have the Entropy Label Used bit set, the Largest LSE Microflow value already reflects any prior hashing of IP fields.

If the Entropy Label Required bit is set and the LSP being set up is using Entropy Label, then the Largest IP Microflow and Largest L4 Microflow SHOULD be omitted. All midpoint LSR SHOULD not look for entropy beyond the Entropy Label.

If the LSP being set up is not using Entropy Label, then the Largest IP Microflow and Largest L4 Microflow SHOULD be included unless the Entropy Label Used bit is set for every contained LSP. The Largest IP Microflow and Largest L4 Microflow SHOULD be omitted if hashing on the IP addresses or IP addresses and ports would yield no greater entropy than hashing on the label stack alone.

3.3. Path Computation Constraints

The RSVP-TE extensions provides a set of requirements to be met by the links which the LSP is to traverse. This set of requirements also serves as the basis for path computation constraints and for admission control constraints.

3.3.1. Link Multipath Capabilities and Path Computation

Three cases are considered. An LSP may have strict ordering constraints. An MPLS-TP LSP is an example of an LSP with strict ordering constraints. This first type of LSP is covered in Section 3.3.1.1. An LSP may have no ordering constraints at all other than the constraint that microflows cannot be reordered. This second case is covered in Section 3.3.1.2. The remaining case is where an LSP has no ordering constraints but carries traffic for other LSP which do have ordering constraints. This third case is covered in Section 3.3.1.3.

3.3.1.1. Path Computation with Ordering Constraints

For an MPLS-TP LSP or other LSP with a strict packet ordering constraint, any link or FA for which the Ordered Aggregate Enabled bit and Entropy Label Multipath are both clear MUST be excluded from the path computation. If the Default to Multipath bit is set on a link, then setting the OA Depth field to one will override that default.

Link with the Ordered Aggregate Enabled bit clear and the Entropy Label Multipath bit set MAY be included in the path computation if the LSR is capable of encapsulating an LSP with a strict packet

ordering constraint with a fixed Entropy Label. If the LSR is not capable of adding an ELI and EL, then these links MUST be excluded from the path computation.

3.3.1.2. Path Computation with No Ordering Constraint

For an MPLS LSP which has no constraint on packet ordering except that microflows must remain in order and does not contain other LSP with ordering constraints, any link for which the Multipath Enabled bit is set can be used. If a link is advertised as a Link Bundle and the Multipath Enabled bit is set for the link, the available bandwidth SHOULD be taken from the "Unreserved Bandwidth" rather than the "Maximum LSP Bandwidth" (see [RFC4201]).

For most LSP, the bandwidth requirement of the largest microflow is not known but an upper bound is known. For example if the LSP aggregates pseudowires or other LSP of no more than some maximum capacity or LSP which have signaled a microflow upper bound, then an upper bound on the largest microflow is known. If this upper bound exceeds the "Maximum LSP Bandwidth" of a given link, then that link MUST be excluded from the path computation.

3.3.1.3. Path Computation for MPLS containing MPLS-TP

To carry LSP which have strict packet ordering requirements and do not have the Entropy Label Used flag set as a client within a server LSP that do not have strict packet ordering requirements, Entropy Label must be added at the server layer LSP to traverse any link or FA that has the Multipath Enabled bit set. For these LSP links which have the Multipath Enabled bit set and the Entropy Label Multipath bit clear MUST be excluded from the path computation.

If the LSR is not capable of adding an Entropy Label, then to carry any client LSP with the Entropy Label Used clear and the OA Depth set to a non-zero value, the server LSP SHOULD exclude any link or FA that has the Multipath Enabled bit set. For these LSP, any link or FA that has the Multipath Enabled bit set and cannot carry a microflow as large as the entire LSP MUST be excluded from the path computation. These LSP MAY be signaled as having strict packet ordering requirements if they can be carried as a single microflow, but this practice is NOT RECOMMENDED.

3.3.2. Link IP Capabilities and Path Computation

An MPLS-TP LSP cannot be reordered. There may be other types of LSP with strict packet ordering requirements. If LSP with strict packet ordering requirements carry IP, using IP headers in the multipath load distribution would violate the packet ordering requirements.

Some LSP cannot be reordered but do not carry IP, and do not carry payloads which could be mistaken as IP. For example, any LSP carrying only pseudowire traffic, where all pseudowires are using a control word carries no payloads which could be mistaken as IP. These type of LSP can be carried within MPLS LSP that allow use of IP header information in multipath load distribution.

This section focuses on Cases in which links or FA must be excluded from path computation based on the settings of the IP related Flags bits in the Multipath Link Capability TLV.

3.3.2.1. LSP without Packet Ordering Requirements

Many LSP carry only IP or predominantly IP, use no hierarchy or have little diversity in the MPLS label stack, and carry far more traffic than can be carried over a single component link in a multipath. Many LSP due to their high capacity, must traverse only multipath which will use IP header information in the multipath traffic distribution.

For these LSP, links must be excluded from the path computation which do not have the IPv4 Enabled Multipath and IPv6 Enabled Multipath bit set (if carrying both IPv4 and IPv6) and do not have either the Default to IP/MPLS Multipath bit set or the IP Optional Multipath bit set.

Hierarchical PSC LSP which require the use IP header information in the multipath traffic distribution MUST NOT set the Ordered Aggregate Enabled bit, MUST set the Default to IP/MPLS Multipath bit, and MUST NOT set the IP Optional Multipath bit in the FA advertisement. The IP Optional Multipath bit MUST be clear because the LSP cannot change the behavior of midpoint LSR, except perhaps in the case of single hop LSP.

3.3.2.2. LSP with Ordering Requirements

The only time that links or FA with both the Ordered Aggregate Enabled bit and the Entropy Label Multipath bit clear can be used is a special case for MPLS-TP LSP that carry only IP traffic. This case does not apply if the MPLS_TP LSP is carrying other LSP or if it is carrying pseudowires.

Where MPLS-TP LSP are carrying only IP, any link or FA with both the Ordered Aggregate Enabled bit and the Entropy Label Multipath bit clear for which the use of IP header information is not disabled or cannot be disabled on a per LSP basis, that link MUST be excluded from the path computation.

Where MPLS-TP LSP are carrying only IP, links MAY be included in the path computation have the IPv4 Enabled Multipath and IPv6 Enabled Multipath bits clear, or have the Default to IP/MPLS Multipath bit clear, or have the IP Optional Multipath bit set. Links with the IP Optional Multipath set, MUST disable use of IP in the load balance for LSP with the IP Multipath Allowed bit clear.

An MPLS-TP LSP are carrying only IP MUST have OA Depth set to one and LSP Min Depth set to one and the IP Multipath Allowed bit cleared. Call admission SHOULD NOT reject an LSP on the basis of OA Depth being set to one if use of IP headers is always disabled, or can be disabled for the specific LSP. If an MPLS-TP is set up this way and then does start to carry other LSP or carry pseudowires, then reordering within the MPLS-TP LSP will occur.

3.3.3. Link Depth Limitations and Path Computation

For any LSP which does not have strict packet ordering constraints, LSP configuration SHOULD include the following parameters.

LSP Min Depth

a minimal acceptable number of label used in multipath traffic distribution,

LSP IP Depth

a minimal label stack depth where the IP header can be used in multipath traffic distribution

For example, if a PSC LSP will carry LSP which in turn carry very high capacity pseudowires using the pseudowire flow label (see [RFC6391]), the flow label is four labels deep. In this case, LSP Min Depth should be four or higher.

For example, if the same PSC LSP will carry LSP which carry IP traffic with no additional labels, then the IP header is two labels deep. In this case, LSP IP Depth should be two or higher.

For an LSP with non-zero LSP Min Depth, all links with Max Depth set to a value below LSP Min Depth MUST be excluded from the LSP Path Computation.

For an LSP with non-zero LSP IP Depth, all links with IP Depth set to a value below LSP IP Depth MUST be excluded from the LSP Path Computation.

If all LSP carried an accurate LSP Min Depth and LSP IP Depth then neither of these parameters would ever have to be configured. In a network with legacy LSR, it may be necessary to configure these

parameters for some LSP. A per-LSP minimum value of each parameter SHOULD be configurable.

4. Backwards Compatibility

Networks today use three forms of multipath.

1. IP ECMP, including IP ECMP at LER using more than one LSP.
2. Ethernet Link Aggregation [IEEE-802.1AX].
3. MPLS Link Bundling [RFC4201].

4.1. Legacy Multipath Behavior

IP ECMP and Ethernet Link Aggregation always distribute traffic over the entire multipath either using information in the MPLS label stack, or using information in a potential IP header, or using both types of information.

One of two behaviors is assumed for link bundles. Either the link bundles place each LSP in its entirety on a single link bundle component link for all LSP, or link bundles distribute traffic over the entire link bundle using the same techniques used for ECMP and Ethernet Link Aggregation. This second behavior is known as the "all ones" component link (see [RFC4201]).

4.2. Networks without Multipath Extensions

Networks exist that are comprised entirely of LSR which do not support these multipath extensions. In these networks there is no way of telling how multipath links will behave. Since an Ethernet Link Aggregation Group (LAG) is advertised as an ordinary link, there is no way to tell that it is a LAG and not an ordinary link.

4.2.1. Networks with MP Capability on all Multipath

Most large core network today rely heavily on the use of multipath. Ethernet Link Aggregation is heavily used and LSR are configured to use the "all ones" component link for all LSP. The "all ones" component link is the default for many Link Bundling implementations used in core networks.

This is equivalent to the following setting in the Multipath Node Capabilities sub-TLV or Multipath Link Capabilities sub-TLV.

1. Clear the Ordered Aggregate Enabled bit and the IP Optional Multipath bit.
2. Set the Multipath Enabled bit, set the Default to Multipath bit, and clear the Entropy Label Multipath bit.
3. If the label stack is used in the multipath traffic distribution, set Max Depth to the number of label stack entries supported, otherwise set it to zero.
4. Since Entropy Label support is not yet widespread, most LSR would behave as if Entropy Label Multipath were clear.
5. If an IP packet under the label stack can be used in the multipath traffic distribution (very common, almost universal in core LSR), set IPv4 Enabled Multipath, set IPv6 Enabled Multipath, set Default to IP/MPLS Multipath, and set IP Depth to the maximum number of label stack entries which can be skipped over before finding the IP stack. Otherwise clear IPv4 Enabled Multipath, clear IPv6 Enabled Multipath and clear Default to IP/MPLS Multipath.
6. On core networks where UDP and TCP ports are rarely used in multipath, clear all UDP and TCP related bits. On networks where multipath is configured to use TCP and UDP port numbers, these bits would be set.

These networks can support very large LSP but cannot support LSP which require strict packet ordering with other labels below such an LSP, such as pseudowire labels. They may also misroute OAM packet which use GAL (see [RFC5586]) if they use the GAL label in determining the load balance. Generally the Link Bundle advertisements indicate a "Maximum LSP Bandwidth" that is equal to the "Unreserved Bandwidth" if Link Bundling is used with the all-ones component link.

Good or bad, if the behavior is consistent, defaults can be configured in other LSR, such that an accurate guess can be made when no Multipath Link Capability TLV is available for a given link.

For example, in many networks, any link of 10 Gb/s or less can be assumed to be a plain link (no multipath) while any link with a capacity greater than 10 Gb/s can be assumed to be a multipath. These assumptions would hold if no 40 Gb/s or 100 Gb/s links are used.

4.2.2. Netowrks with OA Capability on all Multipath

Some networks, particularly edge networks which tend to be lower capacity, do not use Link Aggregation, and if they use Link Bundling at all, configure each LSR to place each LSP in its entirety on a single link bundle component link for all LSP. Some edge equipment only supports this link bundle behavior.

This is equivalent to the following setting in the Multipath Node Capabilities sub-TLV or Multipath Link Capabilities sub-TLV.

Set the Ordered Aggregate Enabled bit,

Clear the Multipath Enabled bit.

All remaining bits in the Flags field should be clear.

The Max Depth and IP Depth should be set to zero.

These networks can support LSP which require strict packet ordering, but cannot support very large LSP.

Like the behavior described in Section 4.2.1, whether this behavior is good or bad, defaults can be configured which accurately guess the capabilities of links for which no Multipath Link Capability TLV is available.

4.2.3. Legacy Netowrks with Mixed MP and OA Links

Some network may support Ethernet Link Aggregation and all or a subset of LSR which place each LSP in its entirety on a single link bundle component link for all LSP.

If the "Maximum LSP Bandwidth" is set as described in Section 4.2.1, then very large LSP can be supported over link bundles. Very large LSP cannot be supported on LSR which place each LSP in its entirety on a single link bundle component link for all LSP, but these are clearly indicated in signaling,

In these mixed networks it may not be possible to reliably support LSP which require strict packet ordering. It is not possible to know where Ethernet Link Aggregation is used and it is not possible to accurately determine Link Bundling behavior on link bundles where "Maximum LSP Bandwidth" is equal to "Unreserved Bandwidth".

Upgrading LSR to support Entropy Label on both LAG and Link Bundles would improve the ability to carry LSP which require strict packet ordering. To gain any benefit the LSP ingress would have to add ELI

and EL.

If not all LSR are upgraded, then the MPLS TE multipath extensions identify those LSR and multipath that have been upgraded.

4.3. Transition to Multipath Extension Support

If a Multipath Node Capability sub-TLV is not advertised (see Section 2.1), then the LSR does not support these multipath extensions. This allows detection of such nodes and if necessary application of defaults to cover legacy multipath such as typical Ethernet Link Aggregation Behavior.

4.3.1. Simple Transitions

For networks with LSR that do not support multipath extensions, transition is easiest if all legacy LSR support and are configured with a common link bundling behavior. If Ethernet Link Aggregation is not used, a single configured default is needed to cover LSR that do not advertise a Multipath Node Capability sub-TLV.

If Ethernet Link Aggregation had been previously used on Legacy LSR, if possible LAG should be disabled and the members of the former LAG configured and advertised as a link bundle which uses the equivalent "all ones" behavior.

If Ethernet Link Aggregation remains but can be identified in some way, such as links with capacity in excess of some value (for example: greater than 10 Gb/s), then defaults can be set up for these LAG. Alternately administrative attributes could be used [RFC3209].

The transition network in this case lacks the ability to determine the largest microflow that can pass through legacy nodes, but this was the case prior to transition for the entire network prior to transition.

4.3.2. More Challenging Transitions

Transition is made more difficult if legacy LSR in a network support Ethernet Link Aggregation but do not support Link Bundle and cannot be identified by simple means, or the newer LSR lack sufficient configuration capability to support conditional defaults.

This situation is most easily handled if a small upgrade to such an LSR can advertise a fixed Multipath Node Capability sub-TLV giving the characteristics of the Ethernet Link Aggregation implementation on that node. Absent of such cooperation, the problem can be solved by configuration on newer LSR which allows association of a Multipath

Node Capability sub-TLV with a specific legacy router ID and possibly a legacy router ID and link.

LSR supporting Multipath Extensions will need to assign default values through configuration to these legacy LSR running Ethernet Link Aggregation. These default values serve to allow LSP which require strict packet ordering to avoid these legacy LSR.

LSR which do not support [RFC4201] may be sufficiently rare that the ability to assign default values per legacy LSR or per [RFC3209] administrative attribute may not be needed in practice.

5. IANA Considerations

[... to be completed ...]

The symbolic constants IANA-TBD-1 through IANA-TBD-3 need to be replaced. Complete instructions, including identification of the number space for each of these will be added to a later version of this internet-draft.

6. Security Considerations

The combination of MPLS, MPLS-TP, and multipath does not introduce any new security threats. The security considerations for MPLS/GMPLS and for MPLS-TP are documented in [RFC5920] and [I-D.ietf-mpls-tp-security-framework].

7. References

7.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC3471] Berger, L., "Generalized Multi-Protocol Label Switching (GMPLS) Signaling Functional Description", RFC 3471, January 2003.
- [RFC4201] Kompella, K., Rekhter, Y., and L. Berger, "Link Bundling in MPLS Traffic Engineering (TE)", RFC 4201, October 2005.
- [RFC5420] Farrel, A., Papadimitriou, D., Vasseur, JP., and A. Ayyangarps, "Encoding of Attributes for MPLS LSP Establishment Using Resource Reservation Protocol Traffic

Engineering (RSVP-TE)", RFC 5420, February 2009.

- [RFC5786] Aggarwal, R. and K. Kompella, "Advertising a Router's Local Addresses in OSPF Traffic Engineering (TE) Extensions", RFC 5786, March 2010.

7.2. Informative References

- [I-D.ietf-mpls-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-06 (work in progress), September 2012.
- [I-D.ietf-mpls-tp-security-framework]
Fang, L., Niven-Jenkins, B., Mansfield, S., and R. Graveman, "MPLS-TP Security Framework", draft-ietf-mpls-tp-security-framework-04 (work in progress), July 2012.
- [I-D.villamizar-mpls-multipath-use]
Villamizar, C., "Use of Multipath with MPLS-TP and MPLS", draft-villamizar-mpls-multipath-use-00 (work in progress), November 2012.
- [IEEE-802.1AX]
IEEE Standards Association, "IEEE Std 802.1AX-2008 IEEE Standard for Local and Metropolitan Area Networks - Link Aggregation", 2006, <<http://standards.ieee.org/getieee802/download/802.1AX-2008.pdf>>.
- [ITU-T.G.800]
ITU-T, "Unified functional architecture of transport networks", 2007, <<http://www.itu.int/rec/T-REC-G/recommendation.asp?parent=T-REC-G.800>>.
- [RFC2475] Blake, S., Black, D., Carlson, M., Davies, E., Wang, Z., and W. Weiss, "An Architecture for Differentiated Services", RFC 2475, December 1998.
- [RFC2702] Awduche, D., Malcolm, J., Agogbua, J., O'Dell, M., and J. McManus, "Requirements for Traffic Engineering Over MPLS", RFC 2702, September 1999.
- [RFC2991] Thaler, D. and C. Hopps, "Multipath Issues in Unicast and Multicast Next-Hop Selection", RFC 2991, November 2000.
- [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.

- [RFC3260] Grossman, D., "New Terminology and Clarifications for Diffserv", RFC 3260, April 2002.
- [RFC3945] Mannie, E., "Generalized Multi-Protocol Label Switching (GMPLS) Architecture", RFC 3945, October 2004.
- [RFC4206] Kompella, K. and Y. Rekhter, "Label Switched Paths (LSP) Hierarchy with Generalized Multi-Protocol Label Switching (GMPLS) Traffic Engineering (TE)", RFC 4206, October 2005.
- [RFC4385] Bryant, S., Swallow, G., Martini, L., and D. McPherson, "Pseudowire Emulation Edge-to-Edge (PWE3) Control Word for Use over an MPLS PSN", RFC 4385, February 2006.
- [RFC5586] Bocci, M., Vigoureux, M., and S. Bryant, "MPLS Generic Associated Channel", RFC 5586, June 2009.
- [RFC5920] Fang, L., "Security Framework for MPLS and GMPLS Networks", RFC 5920, July 2010.
- [RFC6107] Shiimoto, K. and A. Farrel, "Procedures for Dynamically Signaled Hierarchical Label Switched Paths", RFC 6107, February 2011.
- [RFC6391] Bryant, S., Filsfils, C., Drafz, U., Kompella, V., Regan, J., and S. Amante, "Flow-Aware Transport of Pseudowires over an MPLS Packet Switched Network", RFC 6391, November 2011.

Author's Address

Curtis Villamizar (editor)
Outer Cape Cod Network Consulting

Email: curtis@ocnc.com

Network Working Group
Internet-Draft
Intended status: Informational
Expires: August 17, 2013

Y. Weingarten

S. Aldrin
Huawei Technologies
P. Pan
Infinera
J. Ryoo
ETRI
G. Mirsky
Ericsson
February 13, 2013

Requirements for MPLS Shared Mesh Protection
draft-weingarten-mpls-smp-requirements-03.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-TP data paths, in networks that do not employ a control plane for their 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 August 17, 2013.

Copyright Notice

Copyright (c) 2013 IETF Trust and the persons identified as the document authors. All rights reserved.

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

Table of Contents

| | |
|---|----|
| 1. Introduction | 3 |
| 1.1. Protection or Restoration | 4 |
| 1.2. Scope of document | 4 |
| 1.2.1. Relationship to MPLS-TP | 4 |
| 1.3. Contributing Authors | 5 |
| 2. Terminology and Notation | 5 |
| 2.1. Acronyms | 5 |
| 3. SMP Architecture | 5 |
| 3.1. Coordination of resources | 6 |
| 3.2. Control plane or data plane | 7 |
| 4. SMP Network Objectives | 7 |
| 4.1. Configuration and resource reservation | 7 |
| 4.1.1. Checking resource availability | 8 |
| 4.2. Multiple triggers | 8 |
| 4.3. Notification | 9 |
| 4.4. Reversion of protection resources | 9 |
| 4.5. Protection switching time | 10 |
| 4.6. Timers | 10 |
| 4.7. Communicating information and channel | 10 |
| 5. Manageability Considerations | 10 |
| 6. Security Considerations | 11 |
| 7. IANA Considerations | 11 |
| 8. Acknowledgements | 11 |
| 9. Normative References | 11 |
| Authors' Addresses | 12 |

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.

As described in [RFC6372] Shared Mesh Protection (SMP) supports a form of sharing protection resources, while providing protection for multiple data paths that may not have common endpoints and do not share common points of failure. It should be noted that some protection resources may not be shared by multiple protection paths, while other resources are shared. 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 employing 1:1 linear protection by protection paths [APQRE] and [VPQRZ] respectively. The segment [PQR] and all of its protection resources are shared by both of the protection paths.

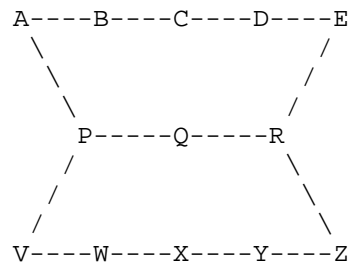


Figure 1: Basic SMP architecture

1.1. Protection or Restoration

[RFC6372], based upon the definitions in [RFC4427], differentiates 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 full allocation of the resources, at least for the shared segments, will only be finalized when 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] 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. 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, Daniel King, Taesik Cheung

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
 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.

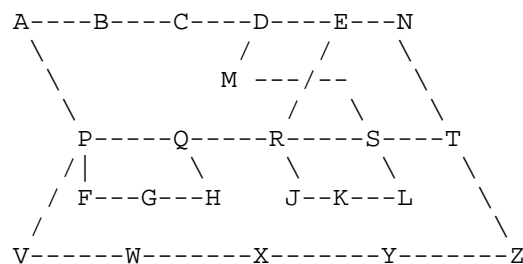


Figure 2: Larger sample SMP architecture

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] is the union of all the shared segments of other protection paths (p1, p2, p3 and p4) while being a whole shared segment of one of the protection paths (p5).

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 group" 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, or shared risk link groups (SRLGs).
- o Determining the size (bandwidth) of the shared resource

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 4.2. 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.

3.2. 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. 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. It should be noted that even when the configuration is performed by a control protocol, e.g. Generalized MPLS (GMPLS), that it is assumed that the control protocol is not used during regular operation of the network.

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. Checking resource availability

When a working path identifies a protection switching trigger it MUST 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. Multiple triggers

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 multiple protection paths of equal priority are requesting allocation of the shared resources, the resources SHOULD be allocated on a first come first served basis. Tie-breaking rules SHALL be defined by the SMP process.
- o If the protection resources are currently in use by a protection path, whose working path has a lower priority, resources required for the higher priority path SHALL be allocated to this path. Traffic with lower priority MAY use available resources or MAY be interrupted.
- o When triggered, protection switching action SHOULD be initiated immediately to minimize service interruption time. If the protection resources are already allocated to a higher priority protection path the protection switching SHALL not be performed.

- o Once a protection path occupies the resource of a shared segments successfully, the traffic on that protection path SHALL NOT be interrupted by any protection traffic whose priority is equal or lower than the protecting path currently in-use.
- o During preemption, shared segment resources MAY be used by both existing traffic (that is being preempted) and higher priority traffic for a short period.
- o During preemption, if there is an oversubscription of resources protected traffic SHOULD be treated as defined in [RFC5712] or [RFC3209]

4.3. Notification

When a working path identifies a trigger for implementing a switchover to the protection path, it SHOULD attempt to switchover the traffic to the protection path and requesting the allocation of the resources for this protected traffic. If the necessary shared resources are in use by a protection path of higher priority or are unavailable to be allocated to the protection path, a notification SHALL be sent to both endpoints of the requesting working path and the switchover MAY not be completed.

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.

4.4. Reversion of protection resources

When the working path detects that the condition that triggered the protection switch has cleared, it is possible to either revert to using the working path resources or continue to utilize the protection resources. Continuing the use of protection resources allows the operator to delay the disruption of service caused by the switchover until periods of lighter traffic. The switchover would need to be performed via an explicit operator command unless the protection resources are preempted by a higher priority fault. The choice between the two modes SHALL depend upon operator configuration. Normally the network should revert to use of the working path resources in order to clear the protection resources for protection of other path triggers. However, the protocol MUST support non-revertive configurations.

4.5. Protection switching time

Protection switching time refers to the transfer time (T_t) defined in [G.808.1] and recovery switching time defined in [RFC4427], and 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 initiate the protection switching process after a failure occurred, and the time needed to complete preemption of existing traffic on the shared segments as described in Section 4.2. The former, which is known as detection and correlation time in [RFC4427] is related to the OAM or management process, but the latter is related to the SMP process. Support for a protection switching time of 50ms is dependent upon the initial switchover to the protection path, but the preemption time SHOULD also be taken into account to minimize total service interruption time.

4.6. Timers

In order to prevent multiple switching actions for a single switching trigger, SMP SHOULD be controlled by 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.

4.7. Communicating information and channel

SMP SHOULD include support for communicating information to coordinate the use of the shared protection resources among multiple working paths. The message encoding and communication channel between the nodes of the shared protection resource and the endpoints of the protection path are out of the scope of this document.

SMP SHOULD provide a communication channel, along the protection path, between the endpoints of the protection path to support fast protection switching.

5. Manageability Considerations

To be added in future version.

6. Security Considerations

To be added in future version.

7. IANA Considerations

This document makes no request of IANA.

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

8. Acknowledgements

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.
- [G.808.1] ITU, "Generic Protection Switching - Linear trail and subnetwork protection", ITU-T G.808.1, Feb 2010.
- [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.

- [RFC5712] Meyer, M. and JP. Vasseur, "MPLS Traffic Engineering Soft Preemption", RFC 5712, January 2010.
- [RFC3209] Awduche, D., Berger, L., Gan, D., Li, T., and V. Srinivasan, "RSVP-TE: Extensions to RSVP for LSP Tunnels", RFC 3209, December 2001.

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

Ping Pan
Infinera

Email: ppan@infinera.com

Jeong-dong Ryoo
ETRI
161 Gajeong
Yuseong, Daejeon 305-700
South Korea

Email: ryoo@etri.re.kr

Greg Mirsky
Ericsson

Email: gregory.mirsky@ericsson.com

