Internet Engineering Task Force
Internet-Draft
Intended status: Standards Track
Expires: August 14, 2014

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

February 10, 2014

Extensions to RSVP-TE for LSP Egress Local Protection
draft-chen-mpls-p2mp-egress-protection-11.txt

Abstract

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

Status of this Memo

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

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

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

This Internet-Draft will expire on August 14, 2014.

Copyright Notice

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

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

Table of Contents

1. Introduction

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

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

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

1.1. An Example of Egress Local Protection

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

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

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

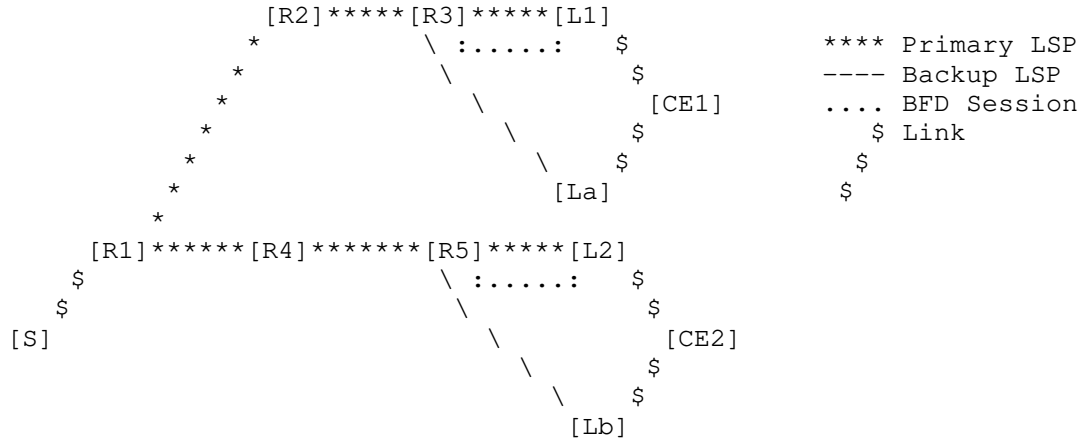


Figure 1: Backup LSP for Locally Protecting Egress

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

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

1.2. Egress Local Protection with FRR

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

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

2. Conventions Used in This Document

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

3. Terminology

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

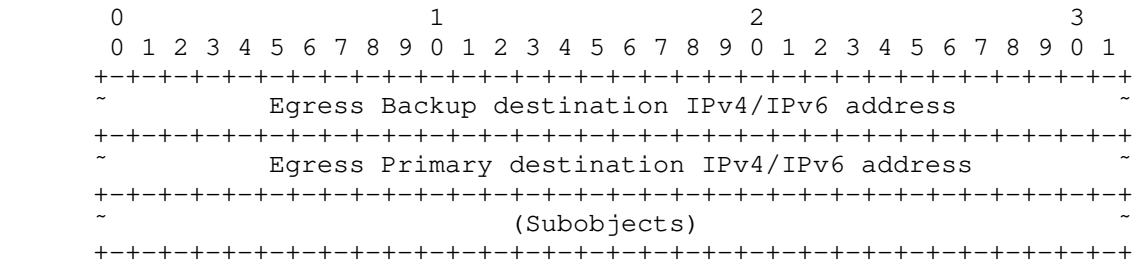
4. Protocol Extensions

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

4.1. EGRESS_BACKUP Object

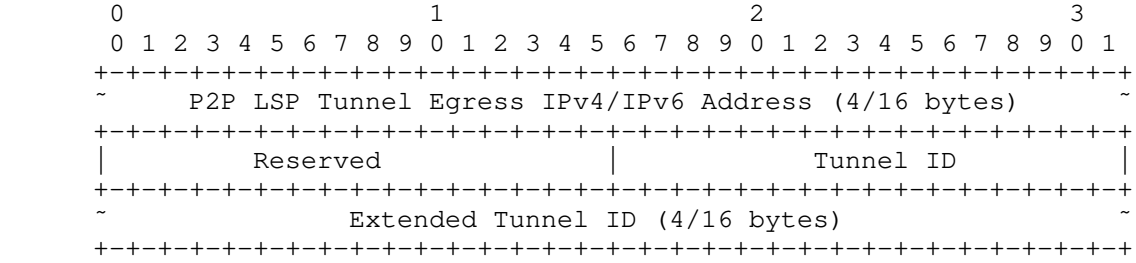
The class of the EGRESS_BACKUP object is TBD-1 to be assigned by IANA. The C-Type of the EGRESS_BACKUP IPv4/IPv6 object is TBD-2/ TBD-3 to be assigned by IANA.

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



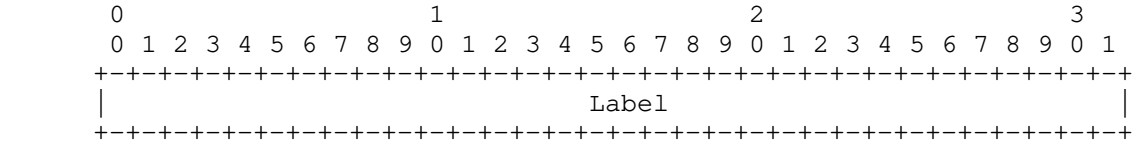
- o Egress Backup destination IPv4/IPv6 address:
  IPv4/IPv6 address of the backup egress node
- o Egress Primary destination IPv4/IPv6 address:
  IPv4/IPv6 address of the primary egress node

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



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

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



4.2. Flags in FAST_REROUTE

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

4.3. Path Message

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

```
<Path Message> ::= <Common Header> [ <INTEGRITY> ]
                   [ [<MESSAGE_ID_ACK> | <MESSAGE_ID_NACK>] ...]
                   [ <MESSAGE_ID> ]<SESSION> <RSVP_HOP> <TIME_VALUES>
                   [ <EXPLICIT_ROUTE> ]
                   <LABEL_REQUEST> [ <PROTECTION> ] [ <LABEL_SET> ...]
                   [ <SESSION_ATTRIBUTE> ] [ <NOTIFY_REQUEST> ]
                   [ <ADMIN_STATUS> ] [ <POLICY_DATA> ... ]
                   <sender descriptor> [<S2L sub-LSP descriptor list>]
                   [<egress backup descriptor list>]
```

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

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

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

5. Egress Protection Behaviors

5.1. Ingress Behavior

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

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

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

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

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

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

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

5.2. Intermediate Node and PLR Behavior

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

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

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

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

After obtaining the backup egress, the PLR tries to compute a path from itself to the backup egress.

The PLR then sets up the backup LSP along the path obtained. It

provides one-to-one backup protection for the primary egress if the "One-to-One Backup Desired" or "One-to-One Backup Preferred" flag is set in the message; otherwise, it provides facility backup protection if the "Facility Backup Desired flag" is set.

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

5.2.1. Signaling for One-to-One Protection

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

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

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

5.2.2. Signaling for Facility Protection

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

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

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

After getting the backup LSP, the PLR associates the backup LSP with a primary LSP for protecting its primary egress. The PLR records

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

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

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

5.2.3. Signaling for S2L Sub LSP Protection

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

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

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

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

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

5.2.4. PLR Procedures during Local Repair

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

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

Moreover, the PLR lets the upstream part of the primary LSP stay after the primary egress fails. The downstream part of the primary LSP from the PLR to the primary egress SHOULD be removed.

6. Considering Application Traffic

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

6.1. A Typical Application

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

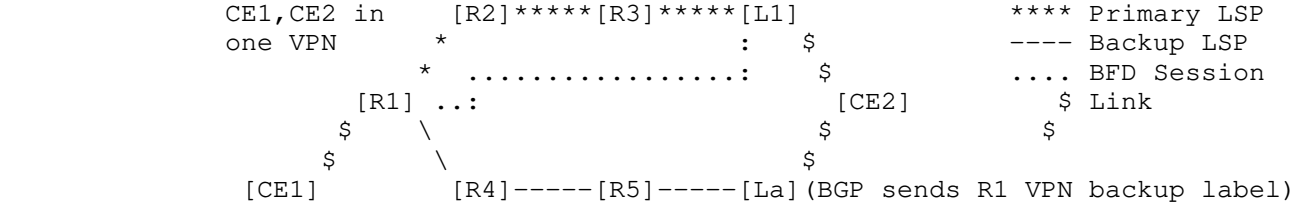


Figure 2: Protect Egress for L3VPN Traffic

In normal operations, R1 sends the traffic from CE1 through primary

LSP with VPN label received from L1 as inner label to L1, which delivers the traffic to CE2 using VPN label.

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

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

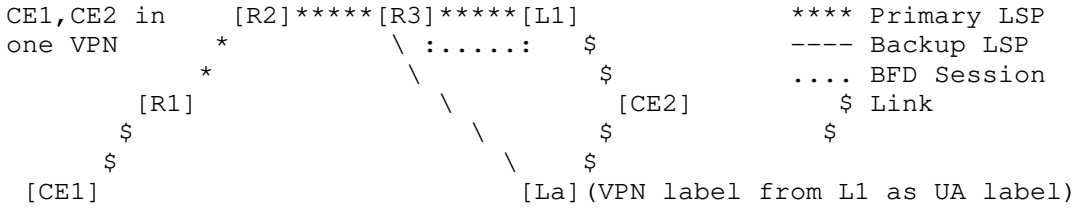


Figure 3: Locally Protect Egress for L3VPN Traffic

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

6.2. PLR Procedure for Applications

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

6.3. Egress Procedures for Applications

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

When the backup egress receives a UA label from the primary egress,

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

7. Security Considerations

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

8. IANA Considerations

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

9. Contributors

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

Zhenbin Li
Huawei Technologies
Huawei Bld., No.156 Beiqing Rd.
Beijing 100095
China
Email: lizhenbin@huawei.com

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

Vic Liu
China Mobile
No.32 Xuanwumen West Street, Xicheng District
Beijing, 100053
China

Email: liuzhiheng@chinamobile.com

10. Acknowledgement

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

11. References

11.1. Normative References

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

[RFC3692] Narten, T., "Assigning Experimental and Testing Numbers Considered Useful", BCP 82, RFC 3692, January 2004.

[RFC2205] Braden, B., Zhang, L., Berson, S., Herzog, S., and S. Jamin, "Resource ReSerVation Protocol (RSVP) -- Version 1 Functional Specification", RFC 2205, September 1997.

[RFC3031] Rosen, E., Viswanathan, A., and R. Callon, "Multiprotocol Label Switching Architecture", RFC 3031, January 2001.

[RFC3209] Awduche, D., Berger, L., Gan, D., Li, T., Srinivasan, V., and G. Swallow, "RSVP-TE: Extensions to RSVP for LSP Tunnels", RFC 3209, December 2001.

[RFC3473] Berger, L., "Generalized Multi-Protocol Label Switching (GMPLS) Signaling Resource ReserVation Protocol-Traffic Engineering (RSVP-TE) Extensions", RFC 3473, January 2003.

[RFC4090] Pan, P., Swallow, G., and A. Atlas, "Fast Reroute Extensions to RSVP-TE for LSP Tunnels", RFC 4090, May 2005.

[RFC4875] Aggarwal, R., Papadimitriou, D., and S. Yasukawa, "Extensions to Resource Reservation Protocol - Traffic Engineering (RSVP-TE) for Point-to-Multipoint TE Label Switched Paths (LSPs)", RFC 4875, May 2007.

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

[RFC5786] Aggarwal, R. and K. Kompella, "Advertising a Router's Local Addresses in OSPF Traffic Engineering (TE) Extensions", RFC 5786, March 2010.

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

11.2. Informative References

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

Authors' Addresses

Huaimo Chen
Huawei Technologies
Boston, MA
USA

Email: huaimo.chen@huawei.com

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

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

Email: fengman.xu@verizon.com

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

Email: mehmet_toy@cable.comcast.com

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

Email: huanglu@chinamobile.com

Lei Liu
UC Davis
USA

Email: liulei.kddi@gmail.com

Internet Engineering Task Force
Internet-Draft
Intended status: Standards Track
Expires: August 18, 2014

H. Chen, Ed.
Huawei Technologies
R. Torvi, Ed.
Juniper Networks

February 14, 2014

Extensions to RSVP-TE for LSP Ingress Local Protection
draft-chen-mpls-p2mp-ingress-protection-11.txt

Abstract

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

Status of this Memo

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

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

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

This Internet-Draft will expire on August 18, 2014.

Copyright Notice

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

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

Table of Contents

1. Co-authors

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

2. Introduction

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

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

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

2.1. An Example of Ingress Local Protection

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

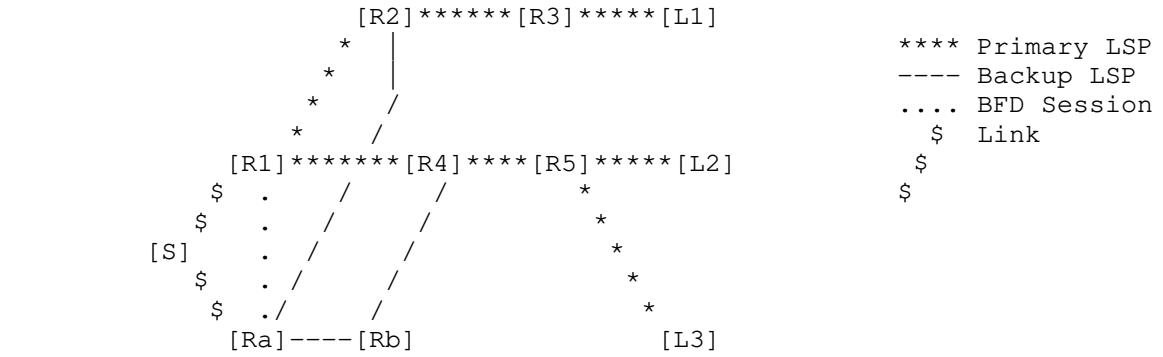


Figure 1: Backup P2MP LSP for Locally Protecting Ingress

Source S may send the traffic simultaneously to both primary ingress R1 and backup ingress Ra. R1 imports the traffic into the primary LSP. Ra normally does not put the traffic into the backup LSP.

Ra should be able to detect the failure of R1 and switch the traffic within 10s of ms. The exact method by which Ra does so is out of scope. Different options are discussed in this draft.

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

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

2.2. Ingress Local Protection with FRR

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

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

3. Ingress Failure Detection

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

3.1. Backup and Source Detect Failure

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

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

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

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

3.2. Backup Detects Failure

Backup Detects Failure or Backup-Detect means that the backup ingress is responsible for detecting the failure of the primary ingress of an LSP. The source SHOULD send the traffic simultaneously to both the primary ingress and backup ingress.

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

Note that the backup ingress may locally distinguish between the failure of the primary ingress and that of the link between the backup ingress and the primary ingress through two BFDs between the backup ingress and the primary ingress. One is through the link, and the other is not. If the first BFD is down and the second is up, the link fails and the primary ingress does not.

3.3. Source Detects Failure

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

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

3.4. Next Hops Detect Failure

Next Hops Detect Failure or Next-Hop-Detect means that each of the next hops of the primary ingress of an LSP is responsible for detecting the failure of the primary ingress.

In normal operations, the source sends the traffic to both the primary ingress and the backup ingress. Both ingresses deliver the traffic to the next hops of the primary ingress. Each of the next

hops selects the traffic from the primary ingress and sends the traffic to the destinations of the LSP.

When each of the next hops detects the failure of the primary ingress, it switches to receive the traffic from the backup ingress and then sends the traffic to the destinations.

3.5. Comparing Different Detection Modes

_Behavior _____ Detection\ Mode	Traffic Always Sent to Backup Ingress	Backup Ingress Activation of Forwarding Entry	Next-Hop Select Stream	Incorrect Failure Detection Cause Traffic Duplication (Ingress does FRR)
Backup-Source-Detect	No	Yes	No	No
Backup-Detect	Yes	Yes	No	Yes
Source-Detect	No	No (Always Active)	No	No
Next-Hop-Detect	Yes	No (Always Active)	Yes	(If Ingress-Next-Hop link fails, stream selection at Next-Next-Hops can mitigate)

A primary goal of failure detection and FRR protection is to avoid traffic duplication, particularly along the P2MP. A reasonable assumption when this ingress protection is in use is that the ingress is also trying to provide link and node protection. When the failure cannot be accurately identified as that of the ingress, this can lead to the ingress sending traffic on bypass to the next-next-hop(s) for node-protection while the backup ingress is sending traffic to its next-hop(s) if Next-Hop-Detect mode is used. RSVP Path messages from the bypass may help to eventually resolve this by removing the forwarding entry for receiving the traffic from the next-hop.

4. Backup Forwarding State

Before the primary ingress fails, the backup ingress is responsible

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

The backup ingress may be off-path or on-path of an LSP. When a backup ingress is not any node of the LSP, we call the backup ingress is off-path. When a backup ingress is a next-hop of the primary ingress of the LSP, we call it is on-path. If the backup ingress is on-path, the primary forwarding state associated with the primary LSP SHOULD be clearly separated from the backup LSP(s) state. Specifically in Backup-Detect mode, the backup ingress will receive traffic from the primary ingress and from the traffic source; only the former should be forwarded until failure is detected even if the backup ingress is the only next-hop.

4.1. Forwarding State for Backup LSP

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

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

4.2. Forwarding State on Next Hops

When Next-Hop-Detect is used, a forwarding entry for a backup LSP is created on each of the next hops of the primary ingress of the LSP. This forwarding entry does not forward any traffic normally. When the primary ingress fails, it is used to import/select the traffic from the backup LSP into the primary LSP.

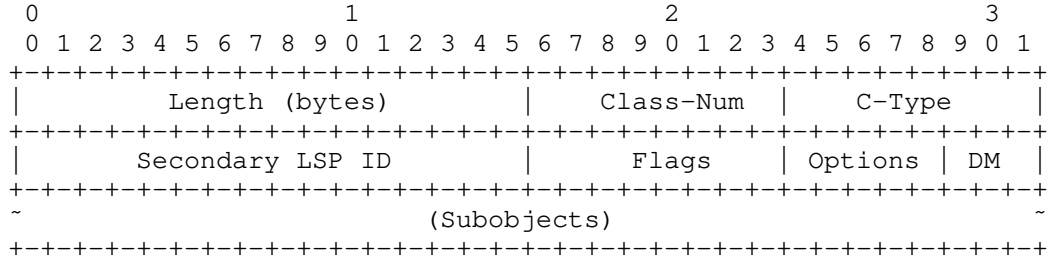
5. Protocol Extensions

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

5.1. INGRESS_PROTECTION Object

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

Class-Num = TBD C-Type = TBD



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

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

DM (Detection Mode)
- 0x00 Backup-Source-Detect
- 0x01 Backup-Detect
- 0x02 Source-Detect
- 0x03 Next-Hop-Detect

For backward compatible, the two high-order bits of the Class-Num in the object are set as follows:

- Class-Num = 0bbbbbbb for the object in a message not on LSP path. The entire message should be rejected and an "Unknown Object Class" error returned.

- Class-Num = 10bbbbbb for the object in a message on LSP path. The node should ignore the object, neither forwarding it nor sending an error message.

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

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

- o Ingress local protection available: The backup ingress sets this flag after backup LSPs are up and ready for locally protecting the primary ingress. The backup ingress sends this to the primary ingress to indicate that the primary ingress is locally protected.

- o Ingress local protection in use: The backup ingress sets this flag when it detects a failure in the primary ingress. The backup ingress keeps it and does not send it to the primary ingress since the primary ingress is down.

- o Bandwidth protection: The backup ingress sets this flag if the backup LSPs guarantee to provide desired bandwidth for the protected LSP against the primary ingress failure.

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

- o Revert to Ingress: The primary ingress sets this option indicating that the traffic for the primary LSP successfully re-signaled will be switched back to the primary ingress from the backup ingress when the primary ingress is restored.

- o Ingress-Proxy/Relay-Message: This option is set to one indicating that Ingress-Proxy method is used. It is set to zero indicating that Relay-Message method is used.

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

The DM (Detection Mode) is used by the primary ingress to specify a desired failure detection mode.

- o Backup-Source-Detect (0x00): The backup ingress and the source are concurrently responsible for detecting the failure involving the primary ingress and redirecting the traffic.

- o Backup-Detect (0x01): The backup ingress is responsible for detecting the failure and redirecting the traffic.

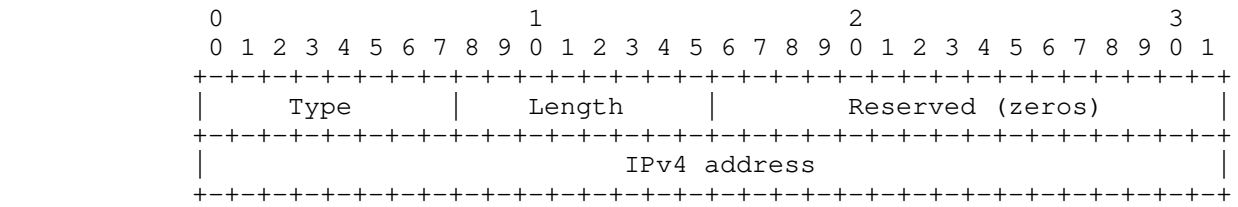
- o Source-Detect (0x02): The source is responsible for detecting the failure and redirecting the traffic.

- o Next-Hop-Detect (0x03): The next hops of the primary ingress are responsible for detecting the failure and selecting the traffic.

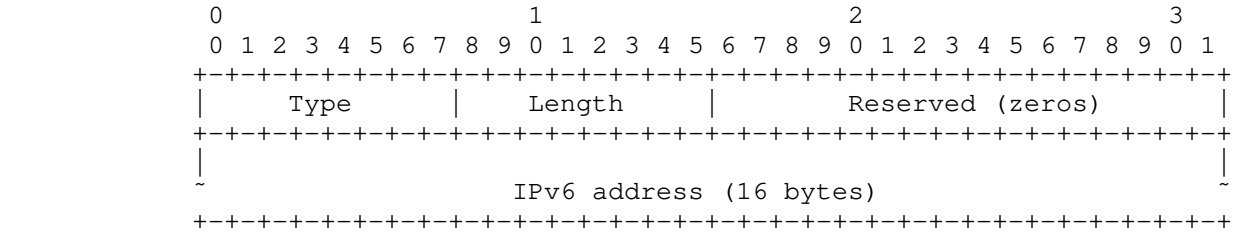
The INGRESS_PROTECTION object may contain some of the sub objects described below.

5.1.1. Subobject: Backup Ingress IPv4/IPv6 Address

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



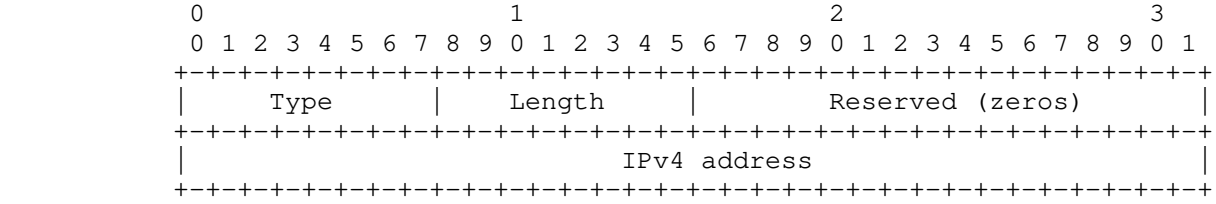
Type: TBD-1 Backup Ingress IPv4 Address
Length: Total length of the subobject in bytes, including the Type and Length fields. The Length is always 8.
Reserved: Reserved two bytes are set to zeros.
IPv4 address: A 32-bit unicast, host address.



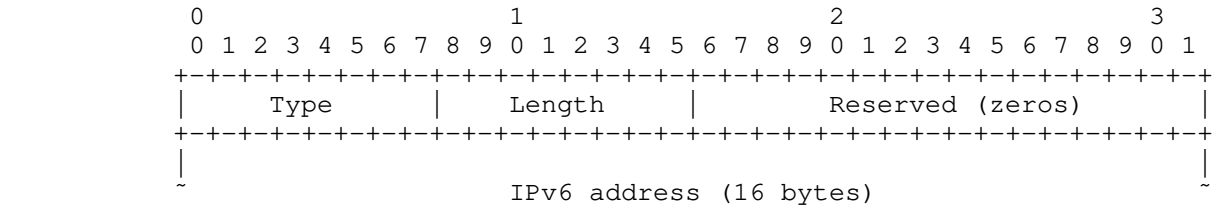
Type: TBD-2 Backup Ingress IPv6 Address
Length: Total length of the subobject in bytes, including the Type and Length fields. The Length is always 20.
Reserved: Reserved two bytes are set to zeros.
IPv6 address: A 128-bit unicast, host address.

5.1.2. Subobject: Ingress IPv4/IPv6 Address

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



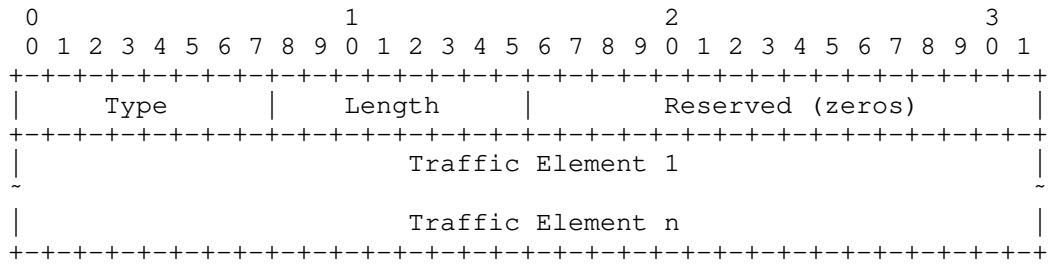
Type: TBD-3 Ingress IPv4 Address
Length: Total length of the subobject in bytes, including the Type and Length fields. The Length is always 8.
Reserved: Reserved two bytes are set to zeros.
IPv4 address: A 32-bit unicast, host address.



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

5.1.3. Subobject: Traffic Descriptor

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



Type: TBD-5/TBD-6/TBD-7 Interface/IPv4/6 Prefix
Length: Total length of the subobject in bytes, including the Type and Length fields.
Reserved: Reserved two bytes are set to zeros.

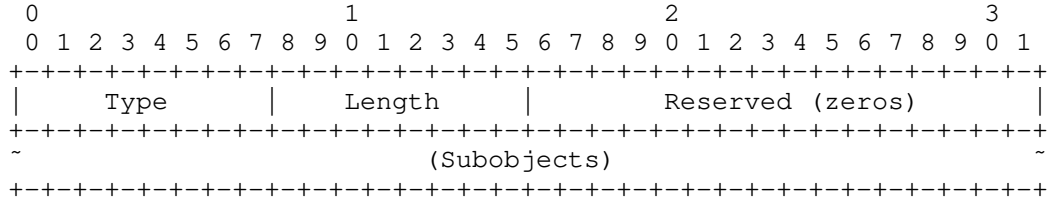
The Traffic Descriptor sub object may contain multiple Traffic Elements of same type as follows.

- Interface Traffic (Type TBD-5): Each of the Traffic Elements is a 32 bit index of an interface, from which the traffic is imported into the backup LSP.

- IPv4/6 Prefix Traffic (Type TBD-6/TBD-7): Each of the Traffic Elements is an IPv4/6 prefix, containing an 8-bit prefix length followed by an IPv4/6 address prefix, whose length, in bits, was specified by the prefix length, padded to a byte boundary.

5.1.4. Subobject: Label-Routes

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



Type: TBD-8 Label-Routes
Length: Total length of the subobject in bytes, including the Type and Length fields.
Reserved: Reserved two bytes are set to zeros.

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

6. Behavior of Ingress Protection

6.1. Overview

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

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

6.1.1. Relay-Message Method

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

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

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

6.1.2. Proxy-Ingress Method

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

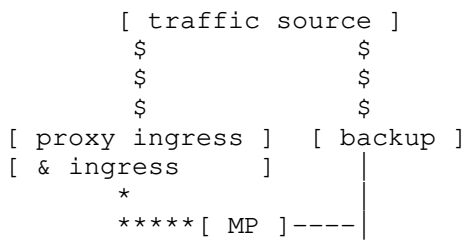


Figure 2: Example Protected LSP with Proxy Ingress Node

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

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

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

6.1.3. Comparing Two Methods

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

6.2. Ingress Behavior

The primary ingress must be configured with four pieces of information for ingress protection.

- Backup Ingress Address: The primary ingress must know an IP address for it to be included in the INGRESS-PROTECTION object.

- Failure Detection Mode: The primary ingress must know what failure detection mode is to be used: Backup-Source-Detect, Backup-Detect, Source-Detect, or Next-Hop-Detect.

- Proxy-Ingress-Id (only needed for Proxy-Ingress Method): The Proxy-Ingress-Id is only used in the Record Route Object for recording the proxy-ingress. If no proxy-ingress-id is specified, then a local interface address that will not otherwise be included in the Record Route Object can be used. A similar technique is used in [RFC4090 Sec 6.1.1].

- Application Traffic Identifier: The primary ingress and backup ingress must both know what application traffic should be directed into the LSP. If a list of prefixes in the Traffic Descriptor sub-object will not suffice, then a commonly understood Application Traffic Identifier can be sent between the primary ingress and backup ingress. The exact meaning of the identifier should be configured similarly at both the primary ingress and backup ingress. The Application Traffic Identifier is understood within the unique context of the primary ingress and backup ingress.

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

6.2.1. Relay-Message Method

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

1. Select a PATH message.

2. If the backup ingress is off-path, then send the backup ingress a PATH message with the content from the selected PATH message and an INGRESS-PROTECTION object; else (the backup ingress is a next hop, i.e., on-path case) add an INGRESS-PROTECTION object into the existing PATH message to the backup ingress (i.e., the next hop). The INGRESS-PROTECTION object contains the Traffic-Descriptor sub-object, the Backup Ingress Address sub-object and the Label-Routes sub-object. The DM (Detection Mode) in the

object is set to indicate the failure detection mode desired. The flags is set to indicate whether a Backup P2MP LSP is desired. If not yet allocated, allocate a second LSP-ID to be used in the INGRESS-PROTECTION object. The Label-Routes sub-object contains the next-hops of the ingress and their labels.

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

6.2.2. Proxy-Ingress Method

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

1. Compute the EROs for the LSP as normal for the ingress.

2. If the selected backup ingress node is not the first node on the path (for all sub-LSPs), then insert at the beginning of the ERO first the backup ingress node and then the ingress node.

3. In the PATH RRO, instead of recording the ingress node's address, replace it with the Proxy-Ingress-Id.

4. Leave the HOP object populated as usual with information for the ingress-node.

5. Add the INGRESS-PROTECTION object to the PATH message. Allocate a second LSP-ID to be used in the INGRESS-PROTECTION object. Include the Backup Ingress Address (IPv4 or IPv6) sub-object and the Traffic-Descriptor sub-object. Set the control-options to indicate the failure detection mode desired. Set or clear the flag indicating that a Backup P2MP LSP is desired.

6. Optionally, add the FAST-REROUTE object [RFC4090] to the Path message. Indicate whether one-to-one backup is desired. Indicate whether facility backup is desired.

7. The RSVP PATH message is sent to the backup node as normal.

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

When the ingress node receives an RSVP PATH message with an INGRESS-PROTECTION object and the object specifies that node as the ingress node and the PHOP as the backup ingress node, the ingress node SHOULD check the Failure Scenario specified in the INGRESS-PROTECTION object and, if it is not the Next-Hop-Detect, then the ingress node SHOULD remove the INGRESS-PROTECTION object from the PATH message before sending it out. Additionally, the ingress node must store that it will install ingress forwarding state for the LSP rather than midpoint forwarding.

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

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

6.3. Backup Ingress Behavior

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

6.3.1. Backup Ingress Behavior in Off-path Case

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

The backup ingress SHOULD follow the control-options specified in the

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

If multiple INGRESS-PROTECTION objects have been received via multiple PATH messages for the same LSP, then the most recent one that specified a Traffic-Descriptor sub-object MUST be the one used.

The backup ingress creates the appropriate forwarding state based on failure detection mode specified. For the Source-Detect and Next-Hop-Detect, this means that the backup ingress forwards any received identified traffic into the backup LSP tunnel(s) to the merge point(s). For the Backup-Detect and Backup-Source-Detect, this means that the backup ingress creates state to quickly determine the primary ingress has failed and switch to sending any received identified traffic into the backup LSP tunnel(s) to the merge point(s).

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

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

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

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

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

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

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

6.3.1.1. Relay-Message Method

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

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

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

6.3.1.2. Proxy-Ingress Method

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

When the backup ingress receives a PATH message with the INGRESS-PROTECTION object, the backup ingress examines the object to learn what traffic associated with the LSP and what ingress failure

detection mode is being used. The backup ingress forwards the PATH message to the ingress node with the normal RSVP changes.

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

6.3.2. Backup Ingress Behavior in On-path Case

An LER as the backup ingress determines that it is on-path if one of its addresses is a next hop of the primary ingress and the primary ingress is not its next hop via checking the PATH message with the INGRESS_PROTECTION object received from the primary ingress. The LER on-path sends the corresponding PATH messages without any INGRESS_PROTECTION object to its next hops. It creates a number of backup P2P LSPs or a backup P2MP LSP from itself to the other next hops (i.e., the next hops other than the backup ingress) of the primary ingress. The other next hops are from the Label-Routes sub object.

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

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

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

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

6.3.3. Failure Detection

Failure detection happens much faster than RSVP, whether via a link-level notification or BFD. As discussed, there are different modes for detecting it. The backup ingress MUST have properly set up its forwarding state to either always forward the specified traffic into the backup LSP(s) for the Source-Detect and Next-Hop-Detect modes or to swap from discarding to forwarding when a failure is detected for the Backup-Source-Detect and Backup-Detect modes.

For facility backup LSPs, the correct inner MPLS label to use must be determined. For the ingress-proxy method, that MPLS label comes directly from the RRO of the RESV. For the relay-message method, that MPLS label comes from the Label-Routes sub-object in the non-empty INGRESS-PROTECTION object.

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

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

6.4. Merge Point Behavior

An LSR that is serving as a Merge Point may need to support the INGRESS-PROTECTION object and functionality defined in this specification if the LSP is ingress-protected where the failure scenario is Next-Hop-Detect. An LSR can determine that it must be a merge point if it is not the ingress, it is not the backup ingress (determined by examining the Backup Ingress Address (IPv4 or IPv6) sub-object in the INGRESS-PROTECTION object), and the PHOP is the ingress node.

In that case, when the LSR receives a PATH message with an INGRESS-PROTECTION object, the LSR MUST remove the INGRESS-PROTECTION object before forwarding on the PATH message. If the failure scenario specified is Next-Hop-Detect, the MP must connect up the fast-failure detection (as configured) to accepting backup traffic received from

the backup node. There are a number of different ways that the MP can enforce not forwarding traffic normally received from the backup node. For instance, first, any LSPs set up from the backup node should not be signaled with an IMPLICIT NULL label and second, the associated label for the ingress- protected LSP could be set to normally discard inside that context.

When the MP receives a RESV message whose matching PATH state had an INGRESS-PROTECTION object, the MP SHOULD add the INGRESS-PROTECTION object to the RESV message before forwarding it. The Backup PATH handling is as described in [RFC4090] and [RFC4875].

6.5. Revertive Behavior

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

- Revert to Primary Ingress: When the primary ingress is restored, it re-signals each of the LSPs that start from the primary ingress. The traffic for every LSP successfully re-signaled is switched back to the primary ingress from the backup ingress.

- Global Repair by Backup Ingress: After determining that the primary ingress of an LSP has failed, the backup ingress computes a new optimal path, signals a new LSP along the new path, and switches the traffic to the new LSP.

6.5.1. Revert to Primary Ingress

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

It is possible that the Ingress failure was inaccurately detected, that the Ingress recovers before the Backup Ingress does Global Repair, or that the Ingress has the ability to take over an LSP based on receiving the associated RESVs.

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

period of time.

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

6.5.2. Global Repair by Backup Ingress

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

When the backup ingress has determined that the primary ingress of the protected LSP has failed (e.g., via the IGP), it can compute a new path and signal a new LSP along the new path so that it no longer relies upon local repair. To do this, the backup ingress uses the same tunnel sender address in the Sender Template Object and uses the previously allocated second LSP-ID in the INGRESS-PROTECTION object of the PATH message as the LSP-ID of the new LSP. This allows the new LSP to share resources with the old LSP. In addition, if the Ingress recovers, the Backup Ingress SHOULD send it RESVs with the INGRESS-PROTECTION object where either the "Force to Backup" or "Revert to Ingress" is specified. The Secondary LSP ID should be the unused LSP ID - while the LSP ID signaled in the RESV will be that currently active. The Ingress can learn from the RESVs what to signal. Even if the Ingress does not take over, the RESVs notify it that the particular LSP IDs are in use. The Backup Ingress can reoptimize the new LSP as necessary until the Ingress recovers. Alternately, the Backup Ingress can create a new LSP with no bandwidth reservation that duplicates the path(s) of the protected LSP, move traffic to the new LSP, delete the protected LSP, and then resignal the new LSP with bandwidth.

7. Security Considerations

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

8. IANA Considerations

TBD

9. Contributors

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

Quintin Zhao
Huawei Technologies
Boston, MA
USA
Email: quintin.zhao@huawei.com

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

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

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

10. Acknowledgement

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

11. References

11.1. Normative References

[RFC1700] Reynolds, J. and J. Postel, "Assigned Numbers", RFC 1700, October 1994.

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

[RFC3692] Narten, T., "Assigning Experimental and Testing Numbers Considered Useful", BCP 82, RFC 3692, January 2004.

[RFC2205] Braden, B., Zhang, L., Berson, S., Herzog, S., and S. Jamin, "Resource ReSerVation Protocol (RSVP) -- Version 1 Functional Specification", RFC 2205, September 1997.

[RFC3031] Rosen, E., Viswanathan, A., and R. Callon, "Multiprotocol Label Switching Architecture", RFC 3031, January 2001.

[RFC3209] Awduche, D., Berger, L., Gan, D., Li, T., Srinivasan, V., and G. Swallow, "RSVP-TE: Extensions to RSVP for LSP Tunnels", RFC 3209, December 2001.

[RFC3473] Berger, L., "Generalized Multi-Protocol Label Switching (GMPLS) Signaling Resource ReserVation Protocol-Traffic Engineering (RSVP-TE) Extensions", RFC 3473, January 2003.

[RFC4090] Pan, P., Swallow, G., and A. Atlas, "Fast Reroute Extensions to RSVP-TE for LSP Tunnels", RFC 4090, May 2005.

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

[RFC4875] Aggarwal, R., Papadimitriou, D., and S. Yasukawa, "Extensions to Resource Reservation Protocol - Traffic Engineering (RSVP-TE) for Point-to-Multipoint TE Label Switched Paths (LSPs)", RFC 4875, May 2007.

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

11.2. Informative References

[RFC2702] Awduche, D., Malcolm, J., Agogbua, J., O'Dell, M., and J. McManus, "Requirements for Traffic Engineering Over MPLS", RFC 2702, September 1999.

[RFC3032] Rosen, E., Tappan, D., Fedorkow, G., Rekhter, Y., Farinacci, D., Li, T., and A. Conta, "MPLS Label Stack Encoding", RFC 3032, January 2001.

Appendix A. Authors' Addresses

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

Ning So
Tata Communications
2613 Fairbourne Cir.
Plano, TX 75082
USA
Email: ning.so@tatacommunications.com

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

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

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

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

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

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

Lei Liu
UC Davis
USA

Email: liulei.kddi@gmail.com

Network Working Group
Internet Draft
Intended status: Informational
Expires: September 14, 2011

Luyuan Fang
Dan Frost
Cisco Systems
Nabil Bitar
Verizon
Raymond Zhang
BT
Lei Wang
Telenor
Kam Lee Yap
XO Communications
Michael Fargano
Qwest
John Drake
Juniper
Thomas Nadeau

March 14, 2011

MPLS-TP OAM Toolset
draft-fang-mpls-tp-oam-toolset-01.txt

Abstract

This document provides an overview of the MPLS-TP OAM toolset, which consists of MPLS-TP fault management and performance monitoring. This overview includes a brief recap of MPLS-TP OAM requirements and functions, and of the generic mechanisms created in the MPLS data plane to support in-band OAM. The importance of using IANA assigned code point under G-Ach when supporting MPLS-TP OAM is also discussed. The protocol definitions for each individual MPLS-TP OAM tool are specified in separate RFCs or Working Group documents which are referenced by this document.

Status of this Memo

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

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

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents

at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress.

This Internet-Draft will expire on September 14, 2011.

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..

Table of Contents

Requirements Language

Although this document is not a protocol specification, the key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [RFC 2119].

1. Introduction

The Operations, Administration, and Maintenance (OAM) Requirements for Transport Profile of Multiprotocol Label Switching (MPLS-TP) networks are defined in RFC 5860 [RFC 5860]. MPLS-TP OAM mechanisms and multiple OAM tools have been developed based on MPLS-TP OAM requirements.

This document provides an overview of the MPLS-TP OAM toolset, which consists of MPLS-TP fault management and performance monitoring. This overview includes a brief recap of MPLS-TP OAM requirements and functions, and of the generic mechanisms created in the MPLS data plane to support in-band OAM. The importance of using IANA assigned code point under G-Ach when supporting MPLS-TP OAM is also discussed.

The protocol definitions for each individual MPLS-TP OAM tool are specified in separate RFCs or Working Group documents while this document is work in progress, which are referenced by this document.

The protocol definitions for each individual MPLS-TP OAM tool are defined in separate RFCs (or Working Group documents while this document is work in progress) referenced by this document.

2. Terminology

This document uses MPLS-TP OAM specific terminology.

Term	Definition
AC	Attachment Circuit
AIS	Alarm indication signal

APS	Automatic Protection Switching
ATM	Asynchronous Transfer Mode
BFD	Bidirectional Forwarding Detection
CC	Continuity Check
CE	Customer-Edge device
CM	Configuration Management
CoS	Class of Service
CV	Connectivity Verification
FM	Fault Management
GAL	Generic Alert Label
G-ACH	Generic Associated Channel
GMPLS	Generalized Multi-Protocol Label Switching
LDI	Link Down Indication
LDP	Label Distribution Protocol
LER	Label Edge Router
LKR	Lock Report
LM	Loss Measurement
LMEG	LSP ME Group
LOC	Loss of Continuity
LSP	Label Switched Path
LSR	Label Switching Router
LSME	LSP SPME ME
LSMEG	LSP SPME ME Group
ME	Maintenance Entity

MEG	Maintenance Entity Group
MEP	Maintenance Entity Group End Point
MIP	Maintenance Entity Group Intermediate Point
MPLS	MultiProtocol Label Switching
NMS	Network Management System
NTP	Network Time Protocol
OAM	Operations, Administration, and Management
PE	Provider Edge
PM	Performance Monitoring
PME	PW Maintenance Entity
PMEG	PW ME Group
PSME	PW SPME ME
PSMEG	PW SPME ME Group
PW	Pseudowire
QoS	Quality of Service
RDI	Remote Defect Indication
SDH	Synchronous Digital Hierarchy
SLA	Service Level Agreement
SME	Section Maintenance Entity
SMEG	Section ME Group
SONET	Synchronous Optical Network
SPME	Sub-path Maintenance Element
S-PE	Switching Provider Edge
SRLG	Shared Risk Link Group
TC	Traffic Class

T-PE Terminating Provider Edge

3. Brief Overview of MPLS-TP OAM Requirements

This following Architectural and Functional Requirements are defined by RFC 5860. They are captured here for easy reading before discussing the toolset.

3.1. Architectural Requirements

The MPLS-TP OAM Supports point-to-point bidirectional PWs, point-to-point co-routed bidirectional LSPs, point-to-point bidirectional Sections, point-to-point associated bidirectional LSPs, point-to-point unidirectional LSPs, and point-to-multipoint LSPs. In addition, MPLS-TP OAM supports these LSPs and PWs when they span single domain or multiple domains.

The protocol solution(s) SHOULD be independent of the underlying tunneling or point-to-point technology or transmission media. The protocol solution(s) SHOULD be independent of the service a PW may emulate.

In-band OAM MUST be implemented. OAM packets for a specific PW, LSP, or Section MUST follow the exact same data path as user traffic of the same.

The solutions MUST support OAM functions with or without relying on IP capabilities.

It is REQUIRED that OAM interoperability be achieved between distinct domains with different operational models, e.g. with IP or without IP support in the data plane.

And OAM functions MUST be configurable even in the absence of a control plane.

3.2. Functional Requirements

In general, MPLS-TP OAM tools MUST provide functions to detect, diagnose, localize, and notify the faults when occur. The mechanism for correction actions trigged by fault detection SHOULD be provided.

The following are the fault detection functional requirements

- Continuity Checks: a function to enable an End Point to monitor the liveness of a PW, LSP, or Section.

- Connectivity Verifications: 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.

- Route Tracing: the functionality to enable an End Point to discover the Intermediate (if any) and End Point(s) along a PW, LSP, or Section, and more generically to trace the route of a PW, LSP or Section.

- Diagnostic Tests: a function to enable conducting diagnostic tests on a PW, LSP, or Section. For example, a loop-back function.

- Lock Instruct: the functionality to enable an End Point of a PW, LSP, or Section to instruct its associated End Point(s) to lock the PW, LSP, or Section.

- Lock Reporting: a function to enable an Intermediate Point of a PW or LSP to report, to an End Point of that same PW or LSP, a lock condition indirectly affecting that PW or LSP.

- Alarm Reporting: the functionality to enable an Intermediate Point of a PW or LSP to report, to an End Point of that same PW or LSP, a fault or defect condition indirectly affecting that PW or LSP.

- Remote Defect Indication: a function to enable an End Point to report, to its associated End Point, a fault or defect condition that it detects on a PW, LSP, or Section for which they are the End Points.

- Client Failure Indication: a function to enable the propagation, from edge to edge of an MPLS-TP network, of information pertaining to a client fault condition detected at an End Point of a PW or LSP, if the client layer OAM does not provide alarm notification.

- Packet Loss Measurement: a function to enable the quantification of packet loss ratio over a PW, LSP, or Section.

- Packet Delay Measurement: a function to enable the quantification of the one-way, and if appropriate, the two-way, delay ratio of a PW, LSP, or Section.

4. MPLS-TP OAM Mechanisms and Toolset Summary

The following subsections provide the summary of MPLS-TP OAM Fault Management and Performance Management toolset, with indication of the corresponding IETF RFCs (or Internet drafts while this document

is work in progress) to support the MPLS-TP OAM functions defined in RFC 5860.

4.1. In-band OAM Mechanisms

To meet the In-band OAM requirements for MPLS-TP, Generic Associated Channel is created [RFC 5586]. It generalizes the applicability of the Pseudowire (PW) Associated Channel Header (ACH) to MPLS Label Switching Paths (LSPs), and Sections.

The Generic Associated Label (GAL) [RFC 5586] is defined by assigning one of the reserved MPLS label values to the G-Ach, GAL identifies the presence of the Associated Channel Header following the label stack.

The creation of G-Ach and GAL provided the necessary mechanisms for building in-band OAM MPLS-TP toolset.

RFC 5718 [RFC 5718] An-In-Band Data Communication Network for the MPLS Transport Profile describes how the G-Ach may be used for Management and Signaling Communication.

4.2. Fault Management Toolset

The following tables provide the summary of MPLS-TP OAM toolset.

Table 1 provides the summary of MPLS-TP OAM Fault Management toolset functions, associated tool/protocol, and the corresponding IETF RFCs or Internet drafts where they are defined.

Table 2 provides the Performance Monitoring Functions, associated tool/protocol definitions, and the corresponding IETF RFCs or Internet Drafts where they are defined.

The following table provide the Performance Monitoring Functions, protocol definitions, and corresponding RFCs or Internet Drafts.

(Editor’s note: only RFCs will be referenced in the final version of the document).

Proactive Fault Management OAM Toolset		
OAM Functions	OAM Tools/Protocols	RFCs / IDs
Continuity Check (CV) & Continuity Verification(CV)	Bidirectional Forwarding Detection (BFD)	draft-ietf-mpls-tp -cc-cv-rdi [cc-cv]
Remote Defect Indication (RDI)	Bidirectional Forwarding Detection (BFD)	draft-ietf-mpls-tp -cc-cv-rdi [cc-cv]
Alarm Indication Signal (AIS)	AIS message under G-Ach	draft-ietf-mpls-tp -fault [fault]
Link Down Indication (LDI)	Flag in AIS message	draft-ietf-mpls-tp -fault [fault]
Lock Report (LKR)	LKR message under G-Ach	draft-ietf-mpls-tp -fault [fault]

Table 1. Proactive Fault Management OAM Toolset

On Demand Fault Management OAM Toolset		
OAM Functions	OAM Tools/Protocols	RFCs / IDs
Continuity Verification(CV)	LSP Ping and BFD	draft-ietf-mpls-tp -cc-cv-rdi [cc-cv]
Diagnostic: Loopback, Lock and LSP Ping	1) In-band Loopback and Lock Instruct 2) LSP Ping	draft-ietf-mpls-tp -li-lb [li-lb]
Lock Instruct (LI)	In-band lock message in G-Ach	draft-ietf-mpls-tp -li-lb [li-lb]

Table 2. On Demand Fault Management OAM Toolset

4.3. Performance Monitoring Toolset

Table 3 provides the Performance Monitoring Fuctions, protocol definitions, and corresponding RFCs or Internet Drafts.

Performance Monitoring OAM Toolset		
OAM Functions	Protocols Definitions	RFCs / IDs
Packet loss measurement (LM)	LM & DM query messages	draft-ietf-mpls-tp -loss-delay [lo-de]
Packet delay (DM) measurement	LM & DM query messages	draft-ietf-mpls-tp -loss-delay -profile [tp-lo-de]
Throughput measurement	derived from Loss measurement	
Delay Variation measurement	Supported from Delay measurement	

Table 3. Performance Monitoring OAM Toolset

5. OAM Toolset Utilization and Protocol Definitions

5.1. Connectivity Check and Connectivity Verification

Continuity Check (CC) and Proactive Connectivity Verification (CV) functions are used to detect loss of continuity (LOC), and unintended connectivity between two MEPs.

Loss of connectivity, mis-merging, mis-connectivity, or unexpected Maintenance Entity Group End Points (MEPs) can be detected using the CC/CV tools.

The CC/CV tools are used to support MPLS-TP fault management, performance management, and protection switching.

Bidirectional Forwarding Detection (BFD) and LSP Ping are defined to support the CC/CV functions [cc-cv].

BFD control packets are sent by the source MEP to sink MEP. The sink MEP monitors the arrival of the BFD control packets and detects the defect.

The interval of BFD control packet can be configured. For example:

- 3.3ms is the default interval for protection switching.
- 100ms is the default interval for performance monitoring.
- 1s is the default interval for fault management.

5.2. Diagnostic Tests and Lock Instruct

The OAM functions to support diagnostic tests are required in the transport environment.

The Loopback mode is defined for management purpose in [li-lb]. The mechanism is provided to Lock and unlock traffic (e.g. data and control traffic) or specific OAM traffic at a specific LSR on the path of the MPLS-TP LSP to allow loop back it to the source by [li-lb].

These diagnostic functions apply to associated bidirectional MPLS-TP LSPs, including MPLS-TP LSPs, bi-directional RSVP-TE tunnels (which is relevant for MPLS-TP dynamic control plane option with GMPLS), and single segment and multi-segment pseudowires.

The Lock operation instruction is carried in an MPLS Loopback request message sent from a MEP to a trail-end MEP of the LSP to request that the LSP be taken out of service. In response, the Lock operation reply is carried in a Loopback response message sent from the trail-end MEP back to the originating MEP to report the result.

The loopback operations include [li-lb]:
- Lock: take an LSP out of service for maintenance.
- Unlock: Restore a previously locked LSP to service.
- Set_Full_Loopback and Set_OAM_Loopback
- Unset_Full_Loopback and Set_OAM_Loopback

Operators can use the loopback mode to test the connectivity or performance (loss, delay, delay variation, and throughput) of given LSP upto a specific node on the path of the LSP.

5.3. Lock Reporting

The Lock Report (LKR) function is used to communicate to the client (sub-) layer MEPs the administrative locking of a server (sub-) layer MEP, and consequential interruption of data traffic forwarding in the client (sub-) layer [fault].

When operator is taking the LSP out of service for maintenance other operational reason, using the LKR function can help to

distinguish the condition as administrative locking from defect condition.

The Lock Report function would also serve the purpose of alarm suppression in the MPLS-TP network above the level of the Lock is occurred. The receipt of an LKR message MAY be treated as the equivalent of loss of continuity at the client layer [fault].

5.4. Alarm Reporting and Link down Indication

Alarm Indication Signal (AIS) message serves the purpose of alarm suppression upon the failure detection in the server (-sub) layer. When the Link Down Indication (RDI) is set, the AIS message MAY be used to trigger recovery mechanisms [fault].

When a server MEP detects the failure, it asserts Loss of Continuity (LOC) or signal fail which sets the flag up to generate OAM packet with AIS message. The AIS message is forwarded to downstream sink MEP in the client layer. This would enable the client layer to suppress the generation of secondary alarms.

A Link Down Indication (LDI) flag is defined in the AIS message. The LDI flag is set in the AIS message in response to detecting a fatal failure in the server layer. Receipt of an AIS message with this flag set MAY be interpreted by a MEP as an indication of signal fail at the client layer. [fault]
Fault OAM messages are generated by intermediate nodes where an LSP is switched, and propagated to the end points (MEPs).

From practical point of view, when both proactive CC functions and LDI are used, one may consider to run the proactive CC functions at a slower rate (e.g. longer BFD hello intervals), and reply on LDI to trigger fast protection switch over upon failure detection in a given LSP.

5.5. Remote Defect Indication

Remote Defect Indication (RDI) function enables an End Point to report to the other End Point that a fault or defect condition is detected on the PW, LSP, or Section they are the End Points.

The RDI OAM function is supported by the use of Bidirectional Forwarding Detection (BFD) Control Packets [cc-cv]. RDI is only used for bidirectional connections and is associated with proactive CC/CV activation.

When an end point (MEP) detects a signal failure condition, it sets the flag up by setting the diagnostic field of the BFD control packet to a particular value to indicate the failure condition on the associated PW, LSP, or Section, and transmitting the BFD control packet with the failure flag up to the other end point (its peer MEP).

RDI function can be used to facilitate the protection switching by synchronizing the two end points when unidirectional failure occurs and is detected by one end.

5.6. Packet Loss and Delay Measurement

Packet loss and delay measurement toolset enables operators to measure the quality of the packet transmission over a PW, LSP, or Section.

The protocol for MPLS-TP loss and delay measurement functions is defined in [lo-de] as profiled in [tp-lo-de]. These documents specify how to measure Packet Loss, Packet Delay, Packet Delay Variation, and Throughput.

The loss and delay protocols have the following characteristics and capabilities:

- Support measurement of packet loss, delay and throughput over Label Switched Paths (LSPs), pseudowires, and MPLS sections (links).

- The same LM and DM protocols can be used for both continuous/proactive and selective/on-demand measurement.

- The LM and DM protocols use a simple query/response model for bidirectional measurement that allows a single node - the querier - to measure the loss or delay in both directions.

- The LM and DM protocols use query messages for unidirectional loss and delay measurement. The measurement can either be carried out at the downstream node(s) or at the querier if an out-of-band return path is available.

- The LM and DM protocols do not require that the transmit and receive interfaces be the same when performing bidirectional measurement.

- The LM protocol supports both 32-bit and 64-bit counters although for simplicity only 32-bit packet counters are currently included in the MPLS-TP profile.

- The LM protocol supports measurement in terms of both packet counts and octet counts although for simplicity only packet counters are currently included in the MPLS-TP profile.

- The LM protocol can be used to measure channel throughput as well as packet loss.

- The DM protocol supports varying the measurement message size in order to measure delays associated with different packet sizes.

6. IANA assigned code points under G-Ach

OAM toolset/functions defined under G-Ach MUST use IANA assigned code points, using Experimental Code Point under G-Ach is inappropriate and it can lead to interoperability problems and potential Code Point collision in production network.

RFC 5586 "MPLS Generic Associated Channel" stated the following in IANA consideration section: A requirement has emerged (see [RFC 5860]) to allow for optimizations or extensions to OAM and other control protocols running in an associated channel to be experimented without resorting to the IETF standards process, by supporting experimental code points. This would prevent code points used for such functions from being used from the range allocated through the IETF standards and thus protects an installed base of equipment from potential inadvertent overloading of code points. In order to support this requirement, IANA has changed the code point allocation scheme for the PW Associated Channel as follows:

0 - 32751: IETF Review
32760 - 32767: Experimental

Code points in the experimental range MUST be used according to the guidelines of RFC 3692 [RFC 3692]. Functions using experimental G-Ach code points MUST be disabled by default.

The guidelines on the usage of experimental numbers are defined in IETF RFC 3692. As indicated by RFC 3692: The experimental numbers are useful when experimenting new protocols or extending existing protocols in order to test and experiment with the new functions, as part of implementation. RFC 3692 reserves a range of numbers for

experimentation when the need of such experimentation has been identified.

However, the experimental numbers "are reserved for generic testing purposes, and other implementations may use the same numbers for different experimental uses." "Experimental numbers are intended for experimentation and testing and are not intended for wide or general deployments." "Shipping a product with a specific value pre-enabled would be inappropriate and can lead to interoperability problems when the chosen value collides with a different usage, as it someday surely will."

Further more, "it would be inappropriate for a group of vendors, a consortia, or another Standards Development Organization to agree among themselves to use a particular value for a specific purpose and then agree to deploy devices using those values." Experimental numbers are not guaranteed to be unique by definition. There is the risk of code point collision when using Experimental Code Point in production networks.

Similar statements can also be found in RFC4929 "Change Process for Multiprotocol Label Switching (MPLS) and Generalized MPLS (GMPLS) Protocols and Procedures". As described in [RFC 4775], "non-standard extensions, including experimental values, are not to be portrayed as industrial standards whether by an individual vendor, an industry forum, or a standards body."

7. Security Considerations

The document provides overview of MPLS-TP OAM requirements, functions, protocol, and solution considerations. The actual protocols for the OAM toolset are defined in separate documents and referenced by this document.

The general security considerations are provided in Security Framework for MPLS and GMPLS Networks [RFC 5920], and MPLS-TP Security Framework [tp-sec-fr].

8. IANA Considerations

This document contains no new IANA considerations.

9. Normative References

[RFC 5586], M. Bocci, M. Vigoureux, S. Bryant, "MPLS Generic Associated Channel",RFC 5586, June 2009.

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

[RFC 5718], D. Beller, and A. Farrel, "An In-Band Data Communication Network For the MPLS Transport Profile", RFC 5718, Jan 2010.

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

[cc-cv] D. Allan, G. Swallow, J. Drake, Proactive Connectivity Verification, Continuity Check and Remote Defect indication for MPLS Transport Profile, draft-ietf-mpls-tp-cc-cv-rdi-03, Feb. 2011.

[fault] G. Swallow, A. Fulignoli, M. Vigoureux, MPLS Fault Management OAM, draft-ietf-mpls-tp-fault-01, March 2011.

[li-lb] S. Boutros, S. Sivabalan, et,al., MPLS Transport Profile Lock Instruct and Loopback Functions draft-ietf-mpls-tp-li-lb-01.txt, March 2011.

[loopback] S. Boutros, S. Sivabalan, G. Swallow, R. Aggarwal, M. Vigoureux, Operating MPLS Transport Profile LSP in Loopback Mode, draft-boutros-mpls-tp-loopback-03.txt, March 2011.

[lo-de] D. Frost, S. Bryant, Packet Loss and Delay Measurement for the MPLS Networks, draft-ietf-mpls-loss-delay-01, Feb. 2011.

[tp-lo-de] D. Frost, S. Bryant, A Packet Loss and Delay Measurement Profile for MPLS-based Transport Networks, draft-frost-mpls-tp-loss-delay-profile-02, Feb. 2011.

10. Informative References

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

[RFC 3692] T. Narten, "Assigning Experimental and Testing Numbers Considered Useful", RFC 3692, Jan. 2004.

[RFC 4775] S. Bradner, "Procedures for Protocol Extensions and Variations", RFC 4775, Dec. 2006.

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

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

[tp-sec-fr] L. Fang, Niven-Jenkins, S. Mansfield, et. Al. MPLS-TP Security Framework, draft-ietf-mpls-tp-security-framework-00, Feb. 2011.

11. Authors' Addresses

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

Dan Frost
Cisco Systems
Email: danfrost@cisco.com

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

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

Lei Wang
Telenor
Telenor Norway
Office Snaroyveien
1331 Fornedbu
Email: Lei.wang@telenor.com

Kam Lee Yap

XO Communications
13865 Sunrise Valley Drive,
Herndon, VA 20171
Email: klyap@xo.com

Michael Fargano
Qwest
5325 Zuni St, 224
Denver CO 80221-1499
Email: Michael.Fargano@qwest.com

John Drake
Juniper
Email: jdrake@juniper.net

Thomas Nadeau
Email: tnadeau@lucidvision.com

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

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

October 25, 2010

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

Abstract

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

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

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

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

Status of this Memo

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

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

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

The list of current Internet-Drafts can be accessed at http://www.ietf.org/ietf/1id-abstracts.txt.

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

This Internet-Draft will expire on January 12, 2011.

Copyright Notice

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

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

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

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

Table of Contents

Requirements Language

Although this document is not a protocol specification, the key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [RFC 2119].

1. Introduction

1.1. Background and Motivation

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

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

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

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

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

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

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

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

1.2. Contributing authors

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

2. Terminologies

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

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

3. Overview of MPLS-TP base functions

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

3.1. MPLS-TP development principles

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

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

3.2. Data Plane

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

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

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

3.3. Control Plane

MPLS-TP allowed two control plane options:

Static: Using NMS for static provisioning;
Dynamic Control Plane using GMPLS, OSPF-TE, RSVP-TE for full automation
ACH concept in PW is extended to GACH for MPLS-TP LSP to support in-band OAM.

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

3.4. OAM

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

1) Continuity Check (CC), Continuity Verification (CV), and Remote Integrity:
- Proactive CC and CV: Extended BFD
- On demand CC and CV: Extended LSP Ping
- Proactive Remote Integrity: Extended BFD
- On demand Remote Integrity: Extended LSP Ping

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

- Static PW defect indication: Use Static PW status

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

3.5. Survivability

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

4. MPLS-TP Use Case Studies

4.1. Mobile Backhaul

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

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

4.1.1. 2G and 3G Mobile Backhaul Support

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

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

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

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

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

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

4.1.2. LTE Mobile Backhaul

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

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

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

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

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

4.1.3. WiMAX Backhaul

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

4.2. Metro Access and Aggregation

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

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

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

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

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

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

4.3. Packet Optical Transport

(to be added)

5. Network Design Considerations

5.1. IP/MPLS vs. MPLS-TP

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

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

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

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

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

5.2. Standards compliance

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

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

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

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

5.3. End-to-end MPLS OAM consistency

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

5.4. Delay and delay variation

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

5.4.1. Network Delay

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

There are three types of network delay:

1. Absolute Delay Time

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

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

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

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

3. Relative Delay Time

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

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

The following are the three technical reasons:

Legacy throughput issue

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

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

Locating Network Acceralators at CE

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

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

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

c) Use of synchronized replication system

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

The following are the network operational feasibility issues.

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

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

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

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

5.5. General network design considerations

- Migration considerations
- Resilency
- Scalability
- Performance

6. MPLS-TP Deployment Consideration

6.1. Network Modes Selection

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

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

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

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

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

Overlay mode

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

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

6.2. Provisioning Modes Selection

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

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

7. Security Considerations

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

8. IANA Considerations

This document contains no new IANA considerations.

9. Normative References

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

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

(More to be added)

10. Informative References

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

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

(More to be added)

11. Author's Addresses

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

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

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

Raymond Zhang
British Telecom

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

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

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

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

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

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

MPLS Working Group
Internet Draft
Updates: 5036, 6720 (if approved)
Intended status: Standards Track
Expires: August 2015

Rajiv Asati
Carlos Pignataro
Kamran Raza
Cisco

Vishwas Manral
Hewlett-Packard, Inc

Rajiv Papneja
Huawei

February 26, 2015

Updates to LDP for IPv6
draft-ietf-mpls-ldp-ipv6-17

Abstract

The Label Distribution Protocol (LDP) specification defines procedures to exchange label bindings over either IPv4, or IPv6 or both networks. This document corrects and clarifies the LDP behavior when IPv6 network is used (with or without IPv4). This document updates RFC 5036 and RFC 6720.

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, 2015.

Copyright Notice

Copyright (c) 2015 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

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

However, RFC5036 specification has the following deficiency (or lacks details) in regards to IPv6 usage (with or without IPv4):

1) LSP Mapping: No rule for mapping a particular packet to a particular LSP that has an Address Prefix FEC element containing IPv6 address of the egress router

2) LDP Identifier: No details specific to IPv6 usage

3) LDP Discovery: No details for using a particular IPv6 destination (multicast) address or the source address

4) LDP Session establishment: No rule for handling both IPv4 and IPv6 transport address optional objects in a Hello message, and subsequently two IPv4 and IPv6 transport connections

5) LDP Address Distribution: No rule for advertising IPv4 or/and IPv6 Address bindings over an LDP session

6) LDP Label Distribution: No rule for advertising IPv4 or/and IPv6 FEC-label bindings over an LDP session, and for handling the co-existence of IPv4 and IPv6 FEC Elements in the same FEC TLV

7) Next Hop Address Resolution: No rule for accommodating the usage of duplicate link-local IPv6 addresses

8) LDP TTL Security: No rule for built-in Generalized TTL Security Mechanism (GTSM) in LDP with IPv6 (this is a deficiency in RFC6720)

This document addresses the above deficiencies by specifying the desired behavior/rules/details for using LDP in IPv6 enabled networks (IPv6-only or Dual-stack networks). This document closes the IPv6 MPLS gap discussed in Sections 3.2.1, 3.2.2, and 3.3.1.1 of [RFC7439].

Note that this document updates RFC5036 and RFC6720.

1.1. Topology Scenarios for Dual-stack Environment

Two LSRs may involve basic and/or extended LDP discovery in IPv6 and/or IPv4 address-families in various topology scenarios.

This document addresses the following 3 topology scenarios in which the LSRs may be connected via one or more Dual-stack LDP enabled interfaces (figure 1), or one or more Single-stack LDP enabled interfaces (figure 2 and figure 3):

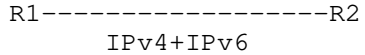


Figure 1 LSRs connected via a Dual-stack Interface

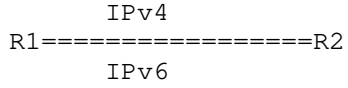


Figure 2 LSRs connected via two Single-stack Interfaces

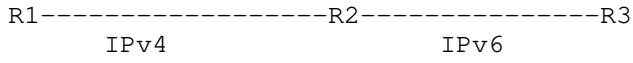


Figure 3 LSRs connected via a Single-stack Interface

Note that the topology scenario illustrated in figure 1 also covers the case of a Single-stack LDP enabled interface (IPv4, say) being converted to a Dual-stacked LDP enabled interface (by enabling IPv6 routing as well as IPv6 LDP), even though the LDPoIPv4 session may already be established between the LSRs.

Note that the topology scenario illustrated in figure 2 also covers the case of two routers getting connected via an additional Single-stack LDP enabled interface (IPv6 routing and IPv6 LDP), even though the LDPoIPv4 session may already be established between the LSRs over the existing interface(s).

This document also addresses the scenario in which the LSRs do the extended discovery in IPv6 and/or IPv4 address-families:

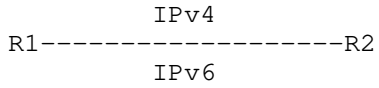


Figure 4 LSRs involving IPv4 and IPv6 address-families

1.2. Single-hop vs. Multi-hop LDP Peering

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

As a consequence, any LDP feature that relies on multi-hop LDP peering session would not work with GTSM and will warrant (statically or dynamically) disabling GTSM. Please see section 10.

2. Specification Language

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

Abbreviations:

LDP - Label Distribution Protocol

LDPoIPv4 - LDP over IPv4 transport connection

LDPoIPv6 - LDP over IPv6 transport connection

FEC - Forwarding Equivalence Class

TLV - Type Length Value

LSR - Label Switching Router

LSP - Label Switched Path

LSPv4 - IPv4-signaled Label Switched Path [RFC4798]

LSPv6 - IPv6-signaled Label Switched Path [RFC4798]

AFI - Address Family Identifier

LDP Id - LDP Identifier

Single-stack LDP - LDP supporting just one address family (for discovery, session setup, address/label binding exchange etc.)

Dual-stack LDP - LDP supporting two address families (for discovery, session setup, address/label binding exchange etc.)

Dual-stack LSR - LSR supporting Dual-stack LDP for a peer

Single-stack LSR - LSR supporting Single-stack LDP for a peer

Note that an LSR can be a Dual-stack and Single-stack LSR at the same time for different peers. This document loosely uses the term address family to mean IP address family.

3. LSP Mapping

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

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

This rule is correct for IPv4, but not for IPv6, since an IPv6 router may even have a /64 or /96 or /128 (or whatever prefix length) address. Hence, that rule is updated to use IPv4 or IPv6 address instead of /32 or /128 addresses as shown below:

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

4. LDP Identifiers

In line with section 2.2.2 of [RFC5036], this document specifies the usage of 32-bit (unsigned non-zero integer) LSR Id on an IPv6 enabled LSR (with or without Dual-stacking).

This document also qualifies the first sentence of last paragraph of Section 2.5.2 of [RFC5036] to be per address family and therefore updates that sentence to the following:

> "For a given address family, an LSR MUST advertise the same transport address in all Hellos that advertise the same label space."

This rightly enables the per-platform label space to be shared between IPv4 and IPv6.

In summary, this document mandates the usage of a common LDP identifier (same LSR Id aka LDP Router Id as well as a common Label space id) for both IPv4 and IPv6 address families.

5. Neighbor Discovery

If Dual-stack LDP is enabled (e.g. LDP enabled in both IPv6 and IPv4 address families) on an interface or for a targeted neighbor, then the LSR MUST transmit both IPv6 and IPv4 LDP (Link or targeted) Hellos and include the same LDP Identifier (assuming per-platform label space usage) in them.

If Single-stack LDP is enabled (e.g. LDP enabled in either IPv6 or IPv4 address family), then the LSR MUST transmit either IPv6 or IPv4 LDP (Link or targeted) Hellos respectively.

5.1. Basic Discovery Mechanism

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

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

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

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

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

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

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

Also, the LDP Link Hello packets MUST have their IPv6 Hop Limit set to 255, be checked for the same upon receipt (before any LDP specific processing) and be handled as specified in Generalized TTL Security Mechanism (GTSM) section 3 of [RFC5082]. The built-in inclusion of GTSM automatically protects IPv6 LDP from off-link attacks.

More importantly, if an interface is a Dual-stack LDP interface (e.g. LDP enabled in both IPv6 and IPv4 address families), then the LSR MUST periodically transmit both IPv6 and IPv4 LDP Link Hellos (using the same LDP Identifier per section 4) on that interface and be able to receive them. This facilitates discovery of IPv6-only, IPv4-only and Dual-stack peers on the interface's subnet and ensures successful subsequent peering using the appropriate (address family) transport on a multi-access or broadcast interface.

5.1.1. Maintaining Hello Adjacencies

In case of Dual-stack LDP enabled interface, the LSR SHOULD maintain link Hello adjacencies for both IPv4 and IPv6 address families. This document, however, allows an LSR to maintain Rx-side Link Hello adjacency only for the address family that has been used for the establishment of the LDP session (whether LDPoIPv4 or LDPoIPv6 session).

5.2. Extended Discovery Mechanism

The extended discovery mechanism (defined in section 2.4.2 of [RFC5036]), in which the targeted LDP Hellos are sent to a unicast IPv6 address destination, requires only one IPv6 specific consideration: the link-local IPv6 addresses MUST NOT be used as the targeted LDP hello packet's source or destination addresses.

6. LDP Session Establishment and Maintenance

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

1. Transport connection establishment
2. Session initialization

The forthcoming sub-section 6.1 discusses the LDP consideration for IPv6 and/or Dual-stacking in the context of session establishment, whereas sub-section 6.2 discusses the LDP consideration for IPv6 and/or Dual-stacking in the context of session maintenance.

6.1. Transport connection establishment

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

This document specifies that:

1. An LSR MUST NOT send a Hello message containing both IPv4 and IPv6 transport address optional objects. In other words, there MUST be at most one optional Transport Address object in a Hello message. An LSR MUST include only the transport address whose address family is the same as that of the IP packet carrying the Hello message.

2. An LSR SHOULD accept the Hello message that contains both IPv4 and IPv6 transport address optional objects, but MUST use only the transport address whose address family is the same as that of the IP packet carrying the Hello message. An LSR SHOULD accept only the first transport object for a given address family in the received Hello message, and ignore the rest, if the LSR receives more than one transport object for a given address family.

3. An LSR MUST send separate Hello messages (each containing either IPv4 or IPv6 transport address optional object) for each IP address family, if Dual-stack LDP is enabled (for an interface or neighbor).

4. An LSR MUST use a global unicast IPv6 address in IPv6 transport address optional object of outgoing targeted Hellos, and check for the same in incoming targeted hellos (i.e. MUST discard the targeted hello, if it failed the check).

5. An LSR MUST prefer using a global unicast IPv6 address in IPv6 transport address optional object of outgoing Link Hellos, if it had to choose between global unicast IPv6 address and unique-local or link-local IPv6 address.

6. A Single-stack LSR MUST establish either LDPoIPv4 or LDPoIPv6 session with a remote LSR as per the enabled address-family.

7. A Dual-stack LSR MUST NOT initiate (or accept the request for) a TCP connection for a new LDP session with a remote LSR, if they already have an LDPoIPv4 or LDPoIPv6 session (for the same LDP Identifier) established.

   This means that only one transport connection is established regardless of IPv6 or/and IPv4 Hello adjacencies presence between two LSRs.

8. A Dual-stack LSR SHOULD prefer establishing an LDPoIPv6 session (instead of LDPoIPv4 session) with a remote Dual-stack LSR by following the 'transport connection role' determination logic in section 6.1.1.

   Additionally, to ensure the above preference in case of Dual-stack LDP being enabled on an interface, it would be desirable that IPv6 LDP Link Hellos are transmitted before IPv4 LDP Link Hellos, particularly when an interface is coming into service or being reconfigured.

6.1.1. Determining Transport connection Roles

Section 2.5.2 of [RFC5036] specifies the rules for determining active/passive roles in setting up TCP connection. These rules are clear for a Single-stack LDP, but not for a Dual-stack LDP, in which an LSR may assume different roles for different address families, causing LDP session to not get established.

To ensure deterministic transport connection (active/passive) role in case of Dual-stack LDP, this document specifies that the Dual-stack LSR conveys its transport connection preference in every LDP Hello message. This preference is encoded in a new TLV, named Dual-stack capability TLV, as defined below:

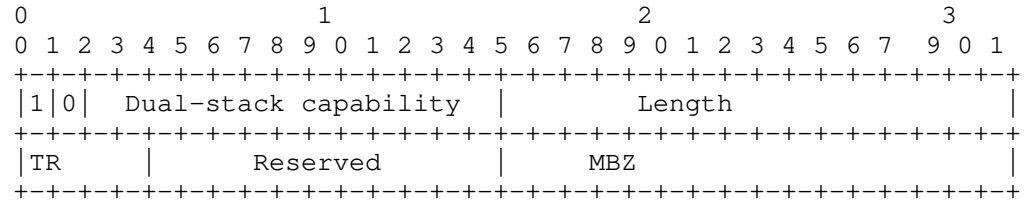


Figure 5 Dual-stack capability TLV

Where:

U and F bits: 1 and 0 (as specified by RFC5036)

Dual-stack capability: TLV code point (to be assigned by IANA).

TR, Transport Connection Preference.

This document defines the following 2 values:

0100: LDPoIPv4 connection

0110: LDPoIPv6 connection (default)

Reserved

This field is reserved. It MUST be set to zero on transmission and ignored on receipt.

A Dual-stack LSR (i.e. LSR supporting Dual-stack LDP for a peer) MUST include "Dual-stack capability" TLV in all of its LDP Hellos, and MUST set the "TR" field to announce its preference for either LDPoIPv4 or LDPoIPv6 transport connection for that peer. The default preference is LDPoIPv6.

A Dual-stack LSR MUST always check for the presence of "Dual-stack capability" TLV in the received hello messages, and take appropriate actions as follows:

1. If "Dual-stack capability" TLV is present and remote preference does not match with the local preference (or does not get recognized), then the LSR MUST discard the hello message and log an error.

   If LDP session was already in place, then LSR MUST send a fatal Notification message with status code [Transport Connection mismatch, IANA allocation TBD] and reset the session.

2. If "Dual-stack capability" TLV is present, and remote preference matches with the local preference, then:

   a) If TR=0100 (LDPoIPv4), then determine the active/passive roles for TCP connection using IPv4 transport address as defined in section 2.5.2 of RFC 5036.

b) If TR=0110 (LDPoIPv6), then determine the active/passive roles for TCP connection by using IPv6 transport address as defined in section 2.5.2 of RFC 5036.

3. If "Dual-stack capability" TLV is NOT present, and

a) Only IPv4 hellos are received, then the neighbor is deemed as a legacy IPv4-only LSR (supporting Single-stack LDP), hence, an LDPoIPv4 session SHOULD be established (similar to that of 2a above).

However, if IPv6 hellos are also received at any time during the life of session from that neighbor, then the neighbor is deemed as a non-compliant Dual-stack LSR (similar to that of 3c below), resulting in any established LDPoIPv4 session being reset and a fatal Notification message being sent (with status code of 'Dual-Stack Non-Compliance', IANA allocation TBD).

b) Only IPv6 hellos are received, then the neighbor is deemed as an IPv6-only LSR (supporting Single-stack LDP) and LDPoIPv6 session SHOULD be established (similar to that of 2b above).

However, if IPv4 hellos are also received at any time during the life of session from that neighbor, then the neighbor is deemed as a non-compliant Dual-stack LSR (similar to that of 3c below), resulting in any established LDPoIPv6 session being reset and a fatal Notification message being sent (with status code of 'Dual-Stack Non-Compliance', IANA allocation TBD).

c) Both IPv4 and IPv6 hellos are received, then the neighbor is deemed as a non-compliant Dual-stack neighbor, and is not allowed to have any LDP session. A Notification message should be sent (with status code of 'Dual-Stack Non-Compliance', IANA allocation TBD).

A Dual-stack LSR MUST convey the same transport connection preference ("TR" field value) in all (link and targeted) Hellos that advertise the same label space to the same peer and/or on same interface. This ensures that two LSRs linked by multiple Hello adjacencies using the same label spaces play the same connection establishment role for each adjacency.

A Dual-stack LSR MUST follow section 2.5.5 of RFC5036 and check for matching Hello messages from the peer (either all Hellos also include the Dual-stack capability (with same TR value) or none do).

A Single-stack LSR do not need to use the Dual-stack capability in hello messages and SHOULD ignore this capability, if received.

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

Note - An alternative to this new Capability TLV could be a new Flag value in LDP Hello message, however, it will get used even in a Single-stack IPv6 LDP networks and linger on forever, even though Dual-stack will not. Hence, this alternative is discarded.

6.2. LDP Sessions Maintenance

This document specifies that two LSRs maintain a single LDP session regardless of number of Link or Targeted Hello adjacencies between them, as described in section 6.1. This is independent of whether:

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

If the last hello adjacency for a given address family goes down (e.g. due to Dual-stack LDP enabled interfaces being converted into a Single-stack LDP enabled interfaces on one LSR etc.), and that address family is the same as the one used in the transport connection, then the transport connection (LDP session) MUST be reset. Otherwise, the LDP session MUST stay intact.

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

7. Binding Distribution

LSRs by definition can be enabled for Dual-stack LDP globally and/or per peer so as to exchange the address and label bindings for both IPv4 and IPv6 address-families, independent of LDPoIPv4 or LDPoIPV6 session between them.

However, there might be some legacy LSRs that are fully RFC 5036 compliant for IPv4, but non-compliant for IPv6 (say, section 3.5.5.1 of RFC 5036), causing them to reset the session upon receiving IPv6 address bindings or IPv6 FEC (Prefix) label bindings from a peer compliant with this document. This is somewhat undesirable, as clarified further Appendix A.1 and A.2.

To help maintain backward compatibility (i.e. accommodate IPv4-only LDP implementations that may not be compliant with RFC 5036 section 3.5.5.1), this specification requires that an LSR MUST NOT send any IPv6 bindings to a peer if peer has been determined as a legacy LSR.

The 'Dual-stack capability' TLV, which is defined in section 6.1.1, is also used to determine if a peer is a legacy (IPv4-only Single-stack) LSR or not.

7.1. Address Distribution

An LSR MUST NOT advertise (via ADDRESS message) any IPv4-mapped IPv6 addresses (defined in section 2.5.5.2 of [RFC4291]), and ignore such addresses, if ever received. Please see Appendix A.3.

If an LSR is enabled with Single-stack LDP for any peer, then it MUST advertise (via ADDRESS message) its local IP addresses as per the enabled address family to that peer, and process received Address messages containing IP addresses as per the enabled address family from that peer.

If an LSR is enabled with Dual-stack LDP for a peer and

1. Is NOT able to find the Dual-stack capability TLV in the incoming IPv4 LDP hello messages from that peer, then the LSR MUST NOT advertise its local IPv6 Addresses to the peer.

2. Is able to find the Dual-stack capability in the incoming IPv4 (or IPv6) LDP Hello messages from that peer, then it MUST advertise (via ADDRESS message) its local IPv4 and IPv6 addresses to that peer.

3. Is NOT able to find the Dual-stack capability in the incoming IPv6 LDP Hello messages, then it MUST advertise (via ADDRESS message) only its local IPv6 addresses to that peer.

   This last point helps to maintain forward compatibility (no need to require this TLV in case of IPv6 Single-stack LDP).

7.2. Label Distribution

An LSR MUST NOT allocate and MUST NOT advertise FEC-Label bindings for link-local or IPv4-mapped IPv6 addresses (defined in section 2.5.5.2 of [RFC4291]), and ignore such bindings, if ever received. Please see Appendix A.3.

If an LSR is enabled with Single-stack LDP for any peer, then it MUST advertise (via Label Mapping message) FEC-Label bindings for the enabled address family to that peer, and process received FEC-Label bindings for the enabled address family from that peer.

If an LSR is enabled with Dual-stack LDP for a peer and

1. Is NOT able to find the Dual-stack capability TLV in the incoming IPv4 LDP hello messages from that peer, then the LSR MUST NOT advertise IPv6 FEC-label bindings to the peer (even if IP capability negotiation for IPv6 address family was done).

2. Is able to find the Dual-stack capability in the incoming IPv4 (or IPv6) LDP Hello messages from that peer, then it MUST advertise FEC-Label bindings for both IPv4 and IPv6 address families to that peer.

3. Is NOT able to find the Dual-stack capability in the incoming IPv6 LDP Hello messages, then it MUST advertise FEC-Label bindings for IPv6 address families to that peer.

   This last point helps to maintain forward compatibility (no need to require this TLV for IPv6 Single-stack LDP).

An LSR MAY further constrain the advertisement of FEC-label bindings for a particular address family by negotiating the IP Capability for a given address family, as specified in [IPPWCap] document. This allows an LSR pair to neither advertise nor receive the undesired FEC-label bindings on a per address family basis to a peer.

If an LSR is configured to change an interface or peer from Single-stack LDP to Dual-stack LDP, then an LSR SHOULD use Typed Wildcard FEC procedures [RFC5918] to request the label bindings for the enabled address family. This helps to relearn the label bindings that may have been discarded before without resetting the session.

8. LDP Identifiers and Duplicate Next Hop Addresses

RFC5036 section 2.7 specifies the logic for mapping the IP routing next-hop (of a given FEC) to an LDP peer so as to find the correct label entry for that FEC. The logic involves using the IP routing next-hop address as an index into the (peer Address) database (which is populated by the Address message containing mapping between each peer's local addresses and its LDP Identifier) to determine the LDP peer.

However, this logic is insufficient to deal with duplicate IPv6 (link-local) next-hop addresses used by two or more peers. The reason is that all interior IPv6 routing protocols (can) use link-local IPv6 addresses as the IP routing next-hops, and 'IPv6 Addressing Architecture [RFC4291]' allows a link-local IPv6 address to be used on more than one links.

Hence, this logic is extended by this specification to use not only the IP routing next-hop address, but also the IP routing next-hop interface to uniquely determine the LDP peer(s). The next-hop address-based LDP peer mapping is to be done through LDP peer address database (populated by Address messages received from the LDP peers), whereas next-hop interface-based LDP peer mapping is to be done through LDP hello adjacency/interface database (populated by hello messages received from the LDP peers).

This extension solves the problem of two or more peers using the same link-local IPv6 address (in other words, duplicate peer addresses) as the IP routing next-hops.

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

9. LDP TTL Security

This document recommends enabling Generalized TTL Security Mechanism (GTSM) for LDP, as specified in [RFC6720], for the LDP/TCP transport connection over IPv6 (i.e. LDPoIPv6). The GTSM inclusion is intended to automatically protect IPv6 LDP peering session from off-link attacks.

[RFC6720] allows for the implementation to statically (configuration) and/or dynamically override the default behavior (enable/disable GTSM) on a per-peer basis. Such a configuration an option could be set on either LSR (since GTSM negotiation would ultimately disable GTSM between LSR and its peer(s)).

LDP Link Hello packets MUST have their IPv6 Hop Limit set to 255, and be checked for the same upon receipt before any further processing, as per section 3 of [RFC5082].

10. IANA Considerations

This document defines a new optional parameter for the LDP Hello Message and two new status codes for the LDP Notification Message.

The 'Dual-Stack capability' parameter requires a code point from the TLV Type Name Space. IANA is requested to allocated a code point from the IETF Consensus range 0x0700-0x07ff for the 'Dual-Stack capability' TLV.

The 'Transport Connection Mismatch' status code requires a code point from the Status Code Name Space. IANA is requested to allocate a code point from the IETF Consensus range and mark the E bit column with a '1'.

The 'Dual-Stack Non-Compliance' status code requires a code point from the Status Code Name Space. IANA is requested to allocate a code point from the IETF Consensus range and mark the E bit column with a '1'.

11. Security Considerations

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

While the security issues relevant for the [RFC5036] are relevant for this document as well, this document reduces the chances of off-link attacks when using IPv6 transport connection by including the use of GTSM procedures [RFC5082]. Please see section 9 for LDP TTL Security details.

Moreover, this document allows the use of IPsec [RFC4301] for IPv6 protection, hence, LDP can benefit from the additional security as specified in [RFC7321] as well as [RFC5920].

12. Acknowledgments

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

Thanks to Bob Thomas for providing critical feedback to improve this document early on.

Many thanks to Eric Rosen, Lizhong Jin, Bin Mo, Mach Chen, Shane Amante, Pranjal Dutta, Mustapha Aissaoui, Matthew Bocci, Mark Tinka, Tom Petch, Kishore Tiruveedhula, Manoj Dutta, Vividh Siddha, Qin Wu, Simon Perreault, Brian E Carpenter, Santosh Esale, Danial Johari and Loa Andersson for thoroughly reviewing this document, and providing insightful comments and multiple improvements.

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

13. Additional Contributors

The following individuals contributed to this document:

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

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

Andre Pelletier
Cisco Systems, Inc.
2000 Innovation Drive
Kanata, ON K2K-3E8, Canada
Email: apelleti@cisco.com

14. References

14.1. Normative References

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

[RFC4291] Hinden, R. and S. Deering, "Internet Protocol Version 6 (IPv6) Addressing Architecture", RFC 4291, February 2006.

[RFC5036] Andersson, L., Minei, I., and Thomas, B., "LDP Specification", RFC 5036, October 2007.

[RFC5082] Pignataro, C., Gill, V., Heasley, J., Meyer, D., and Savola, P., "The Generalized TTL Security Mechanism (GTSM)", RFC 5082, October 2007.

[RFC5918] Asati, R., Minei, I., and Thomas, B., "Label Distribution Protocol (LDP) 'Typed Wildcard Forward Equivalence Class (FEC)", RFC 5918, October 2010.

14.2. Informative References

[RFC4301] Kent, S. and K. Seo, "Security Architecture and Internet Protocol", RFC 4301, December 2005.

[RFC7321] Manral, V., "Cryptographic Algorithm Implementation Requirements for Encapsulating Security Payload (ESP) and Authentication Header (AH)", RFC 7321, April 2007.

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

[RFC4798] De Clercq, et al., "Connecting IPv6 Islands over IPv4 MPLS Using IPv6 Provider Edge Routers (6PE)", RFC 4798, February 2007.

[IPPWCap] Raza, K., "LDP IP and PW Capability", draft-ietf-mpls-ldp-ip-pw-capability, October 2014.

[RFC5340] Coltun, R., Ferguson, D., Moy, J., and A. Lindem, "OSPF for IPv6", RFC 5340, July 2008.

[RFC6286] E. Chen, and J. Yuan, "Autonomous-System-Wide Unique BGP Identifier for BGP-4", RFC 6286, June 2011.

[RFC6720] R. Asati, and C. Pignataro, "The Generalized TTL Security Mechanism (GTSM) for the Label Distribution Protocol (LDP)", RFC 6720, August 2012.

[RFC4038] M-K. Shin, Y-G. Hong, J. Hagino, P. Savola, and E. M. Castro, "Application Aspects of IPv6 Transition", RFC 4038, March 2005.

[RFC7439] W. George, and C. Pignataro, "Gap Analysis for Operating IPv6-Only MPLS Networks", RFC 7439, January 2015.

Appendix A.

A.1. LDPv6 and LDPv4 Interoperability Safety Net

It is not safe to assume that RFC5036 compliant implementations have supported handling IPv6 address family (IPv6 FEC label) in Label Mapping message all along.

If a router upgraded with this specification advertised both IPv4 and IPv6 FECs in the same label mapping message, then an IPv4-only peer (not knowing how to process such a message) may abort processing the entire label mapping message (thereby discarding even the IPv4 label FECs), as per the section 3.4.1.1 of RFC5036.

This would result in LDPv6 to be somewhat undeployable in existing production networks.

The change proposed in section 7 of this document provides a good safety net and makes LDPv6 incrementally deployable without making any such assumption on the routers' support for IPv6 FEC processing in current production networks.

A.2. Accommodating Non-RFC5036-compliant implementations

It is not safe to assume that implementations have been RFC5036 compliant in gracefully handling IPv6 address family (IPv6 Address List TLV) in Address message all along.

If a router upgraded with this specification advertised IPv6 addresses (with or without IPv4 addresses) in Address message, then an IPv4-only peer (not knowing how to process such a message) may not follow section 3.5.5.1 of RFC5036, and tear down the LDP session.

This would result in LDPv6 to be somewhat undeployable in existing production networks.

The changes proposed in section 6 and 7 of this document provides a good safety net and makes LDPv6 incrementally deployable without making any such assumption on the routers' support for IPv6 FEC processing in current production networks.

A.3. Why prohibit IPv4-mapped IPv6 addresses in LDP

Per discussion with 6MAN and V6OPS working groups, the overwhelming consensus was to not promote IPv4-mapped IPv6 addresses appear in the routing table, as well as in LDP (address and label) databases.

Also, [RFC4038] section 4.2 suggests that IPv4-mapped IPv6 addressed packets should never appear on the wire.

A.4. Why 32-bit value even for IPv6 LDP Router ID

The first four octets of the LDP identifier, the 32-bit LSR Id (e.g. (i.e. LDP Router Id), identify the LSR and is a globally unique value within the MPLS network. This is regardless of the address family used for the LDP session.

Please note that 32-bit LSR Id value would not map to any IPv4-address in an IPv6 only LSR (i.e., single stack), nor would there be an expectation of it being IP routable, nor DNS-resolvable. In IPv4 deployments, the LSR Id is typically derived from an IPv4 address, generally assigned to a loopback interface. In IPv6 only deployments, this 32-bit LSR Id must be derived by some other means that guarantees global uniqueness within the MPLS network, similar to that of BGP Identifier [RFC6286] and OSPF router ID [RFC5340].

This document reserves 0.0.0.0 as the LSR Id, and prohibits its usage with IPv6, in line with OSPF router Id in OSPF version 3 [RFC5340].

Author's Addresses

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

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

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

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

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

MPLS
Internet-Draft
Intended status: Standards Track
Expires: January 20, 2012

D. Frost
S. Bryant
Cisco Systems

July 19, 2011

Packet Loss and Delay Measurement for MPLS Networks
draft-ietf-mpls-loss-delay-04

Abstract

Many service provider service level agreements (SLAs) depend on the ability to measure and monitor performance metrics for packet loss and one-way and two-way delay, as well as related metrics such as delay variation and channel throughput. This measurement capability also provides operators with greater visibility into the performance characteristics of their networks, thereby facilitating planning, troubleshooting, and evaluation. This document specifies protocol mechanisms to enable the efficient and accurate measurement of these performance metrics in MPLS networks.

Requirements Language

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

Status of this Memo

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

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

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

This Internet-Draft will expire on January 20, 2012.

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.

Table of Contents

1. Introduction

Many service provider service level agreements (SLAs) depend on the ability to measure and monitor performance metrics for packet loss and one-way and two-way delay, as well as related metrics such as delay variation and channel throughput. This measurement capability also provides operators with greater visibility into the performance characteristics of their networks, thereby facilitating planning, troubleshooting, and evaluation. This document specifies protocol mechanisms to enable the efficient and accurate measurement of these performance metrics in MPLS networks.

This document specifies two closely-related protocols, one for packet loss measurement (LM) and one for packet delay measurement (DM). These protocols have the following characteristics and capabilities:

- The LM and DM protocols are intended to be simple and to support efficient hardware processing.

- The LM and DM protocols operate over the MPLS Generic Associated Channel (G-ACh) [RFC5586] and support measurement of loss, delay, and related metrics over Label Switched Paths (LSPs), pseudowires, and MPLS sections (links).

- The LM and DM protocols are applicable to the LSPs, pseudowires, and sections of networks based on the MPLS Transport Profile (MPLS-TP), because the MPLS-TP is based on a standard MPLS data plane. The MPLS-TP is defined and described in [RFC5921], and MPLS-TP LSPs, pseudowires, and sections are discussed in detail in [RFC5960]. A profile describing the minimal functional subset of the LM and DM protocols in the MPLS-TP context is provided in [I-D.ietf-mpls-tp-loss-delay-profile].

- The LM and DM protocols can be used both for continuous/proactive and selective/on-demand measurement.

- The LM and DM protocols use a simple query/response model for bidirectional measurement that allows a single node - the querier - to measure the loss or delay in both directions.

- The LM and DM protocols use query messages for unidirectional loss and delay measurement. The measurement can either be carried out at the downstream node(s) or at the querier if an out-of-band return path is available.

- The LM and DM protocols do not require that the transmit and receive interfaces be the same when performing bidirectional measurement.

- The DM protocol is stateless.

- The LM protocol is "almost" stateless: loss is computed as a delta between successive messages, and thus the data associated with the last message received must be retained.

- The LM protocol can perform two distinct kinds of loss measurement: it can measure the loss of specially generated test messages in order to infer the approximate data-plane loss level (inferred measurement); or it can directly measure data-plane packet loss (direct measurement). Direct measurement provides perfect loss accounting, but may require specialized hardware support and is only applicable to some LSP types. Inferred measurement provides only approximate loss accounting but is generally applicable.

  The direct LM method is also known as "frame-based" in the context of Ethernet transport networks [Y.1731]. Inferred LM is a generalization of the "synthetic" measurement approach currently in development for Ethernet networks, in the sense that it allows test messages to be decoupled from measurement messages.

- The LM protocol supports measurement in terms of both packet counts and octet counts.

- The LM protocol supports both 32-bit and 64-bit counters.

- The LM protocol can be used to measure channel throughput as well as packet loss.

- The DM protocol supports multiple timestamp formats, and provides a simple means for the two endpoints of a bidirectional connection to agree on a preferred format. This procedure reduces to a triviality for implementations supporting only a single timestamp format.

- The DM protocol supports varying the measurement message size in order to measure delays associated with different packet sizes.

The One-Way Active Measurement Protocol (OWAMP) [RFC4656] and Two-Way Active Measurement Protocol (TWAMP) [RFC5357] provide capabilities for the measurement of various performance metrics in IP networks. These protocols are not streamlined for hardware processing and rely on IP and TCP, as well as elements of the Network Time Protocol (NTP), which may not be available or optimized in some network environments; they also lack support for IEEE 1588 timestamps and direct-mode LM, which in some environments may be required. The protocols defined in this document thus are similar in some respects

to, but also differ from, these IP-based protocols.

1.1. Applicability and Scope

This document specifies measurement procedures and protocol messages that are intended to be applicable in a wide variety of circumstances, and amenable to implementation by a wide range of hardware- and software-based measurement systems. As such, it does not attempt to mandate measurement quality levels or analyze specific end-user applications.

1.2. Terminology

Term	Definition
ACH	Associated Channel Header
DM	Delay Measurement
ECMP	Equal Cost Multipath
G-ACh	Generic Associated Channel
LM	Loss Measurement
LSE	Label Stack Entry
LSP	Label Switched Path
NTP	Network Time Protocol
OAM	Operations, Administration, and Maintenance
PTP	Precision Time Protocol
TC	Traffic Class

2. Overview

This section begins with a summary of the basic methods used for the bidirectional measurement of packet loss and delay. These measurement methods are then described in detail. Finally a list of practical considerations are discussed that may come into play to inform or modify these simple procedures. This section is limited to theoretical discussion; for protocol specifics the reader is referred to Section 3 and Section 4.

2.1. Basic Bidirectional Measurement

The following figure shows the reference scenario.

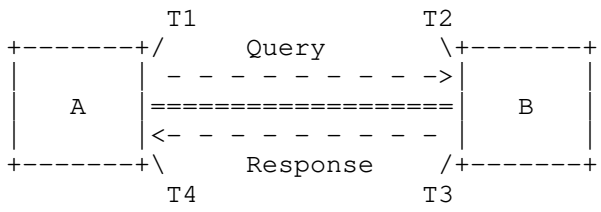


Figure 1

The figure shows a bidirectional channel between two nodes, A and B, and illustrates the temporal reference points T1-T4 associated with a measurement operation that takes place at A. The operation consists of A sending a query message to B, and B sending back a response. Each reference point indicates the point in time at which either the query or the response message is transmitted or received over the channel.

In this situation, A can arrange to measure the packet loss over the channel in the forward and reverse directions by sending Loss Measurement (LM) query messages to B each of which contains the count of packets transmitted prior to time T1 over the channel to B (A_TxP). When the message reaches B, it appends two values and reflects the message back to A: the count of packets received prior to time T2 over the channel from A (B_RxP), and the count of packets transmitted prior to time T3 over the channel to A (B_TxP). When the response reaches A, it appends a fourth value, the count of packets received prior to time T4 over the channel from B (A_RxP).

These four counter values enable A to compute the desired loss statistics. Because the transmit count at A and the receive count at B (and vice versa) may not be synchronized at the time of the first message, and to limit the effects of counter wrap, the loss is computed in the form of a delta between messages.

To measure at A the delay over the channel to B, a Delay Measurement (DM) query message is sent from A to B containing a timestamp recording the instant at which it is transmitted, i.e. T1. When the message reaches B, a timestamp is added recording the instant at which it is received (T2). The message can now be reflected from B to A, with B adding its transmit timestamp (T3) and A adding its receive timestamp (T4). These four timestamps enable A to compute the one-way delay in each direction, as well as the two-way delay for the channel. The one-way delay computations require that the clocks of A and B be synchronized; mechanisms for clock synchronization are outside the scope of this document.

2.2. Packet Loss Measurement

Suppose a bidirectional channel exists between the nodes A and B. The objective is to measure at A the following two quantities associated with the channel:

A_TxLoss (transmit loss): the number of packets transmitted by A over the channel but not received at B;

A_RxLoss (receive loss): the number of packets transmitted by B over the channel but not received at A.

This is accomplished by initiating a Loss Measurement (LM) operation at A, which consists of transmission of a sequence of LM query messages (LM[1], LM[2], ...) over the channel at a specified rate, such as one every 100 milliseconds. Each message LM[n] contains the following value:

A_TxP[n]: the total count of packets transmitted by A over the channel prior to the time this message is transmitted.

When such a message is received at B, the following value is recorded in the message:

B_RxP[n]: the total count of packets received by B over the channel at the time this message is received (excluding the message itself).

At this point, B transmits the message back to A, recording within it the following value:

B_TxP[n]: the total count of packets transmitted by B over the channel prior to the time this response is transmitted.

When the message response is received back at A, the following value is recorded in the message:

A_RxP[n]: the total count of packets received by A over the channel at the time this response is received (excluding the message itself).

The transmit loss A_TxLoss[n-1,n] and receive loss A_RxLoss[n-1,n] within the measurement interval marked by the messages LM[n-1] and LM[n] are computed by A as follows:

A_TxLoss[n-1,n] = (A_TxP[n] - A_TxP[n-1]) - (B_RxP[n] - B_RxP[n-1])
A_RxLoss[n-1,n] = (B_TxP[n] - B_TxP[n-1]) - (A_RxP[n] - A_RxP[n-1])

where the arithmetic is modulo the counter size.

(Strictly speaking, it is not necessary that the fourth count, A_RxP[n], actually be written in the message, but this is convenient for some implementations and useful if the message is to be forwarded on to an external measurement system.)

The derived values

```
A_TxLoss = A_TxLoss[1,2] + A_TxLoss[2,3] + ...

A_RxLoss = A_RxLoss[1,2] + A_RxLoss[2,3] + ...
```

are updated each time a response to an LM message is received and processed, and represent the total transmit and receive loss over the channel since the LM operation was initiated.

When computing the values A_TxLoss[n-1,n] and A_RxLoss[n-1,n] the possibility of counter wrap must be taken into account. Consider for example the values of the A_TxP counter at sequence numbers n-1 and n. Clearly if A_TxP[n] is allowed to wrap to 0 and then beyond to a value equal to or greater than A_TxP[n-1], the computation of an unambiguous A_TxLoss[n-1,n] value will be impossible. Therefore the LM message rate MUST be sufficiently high, given the counter size and the speed and minimum packet size of the underlying channel, that this condition cannot arise. For example, a 32-bit counter for a 100 Gbps link with a minimum packet size of 64 bytes can wrap in 2^32 / (10^11/(64*8)) = ~22 seconds, which is therefore an upper bound on the LM message interval under such conditions. This bound will be referred to as the MaxLMInterval of the channel. It is clear that the MaxLMInterval will be a more restrictive constraint in the case of direct LM and for smaller counter sizes.

The loss measurement approach described in this section has the characteristic of being stateless at B and "almost" stateless at A. Specifically, A must retain the data associated with the last LM response received, in order to use it to compute loss when the next response arrives. This data MAY be discarded, and MUST NOT be used as a basis for measurement, if MaxLMInterval elapses before the next response arrives, because in this case an unambiguous measurement cannot be made.

The foregoing discussion has assumed the counted objects are packets, but this need not be the case. In particular, octets may be counted instead. This will, of course, reduce the MaxLMInterval accordingly.

In addition to absolute aggregate loss counts, the individual loss counts yield additional metrics such as the average loss rate over

any multiple of the measurement interval. An accurate loss rate can be determined over time even in the presence of anomalies affecting individual measurements, such as those due to packet misordering (Section 4.2.10).

Note that an approach for conducting packet loss measurement in IP networks is documented in [RFC2680]. This approach differs from the one described here, for example by requiring clock synchronization between the measurement points and lacking support for direct-mode LM.

2.3. Throughput Measurement

If LM query messages contain a timestamp recording their time of transmission, this data can be combined with the packet or octet counts to yield measurements of the throughput offered and delivered over the channel during the interval in terms of the counted units.

For a bidirectional channel, for example, given any two LM response messages (separated in time by not more than the MaxLMInterval), the difference between the counter values tells the querier the number of units successfully transmitted and received in the interval between the timestamps. Absolute offered throughput is the number of data units transmitted and absolute delivered throughput is the number of data units received. Throughput rate is the number of data units sent or received per unit time.

Just as for loss measurement, the interval counts can be accumulated to arrive at the absolute throughput of the channel since the start of the measurement operation, or used to derive related metrics such as the throughput rate. This procedure also enables out-of-service throughput testing when combined with a simple packet generator.

2.4. Delay Measurement

Suppose a bidirectional channel exists between the nodes A and B. The objective is to measure at A one or more of the following quantities associated with the channel:

- The one-way delay associated with the forward (A to B) direction of the channel;
- The one-way delay associated with the reverse (B to A) direction of the channel;
- The two-way delay (A to B to A) associated with the channel.

The one-way delay metric for packet networks is described in

[RFC2679]. In the case of two-way delay, there are actually two possible metrics of interest. The "two-way channel delay" is the sum of the one-way delays in each direction and reflects the delay of the channel itself, irrespective of processing delays within the remote endpoint B. The "round-trip delay" is described in [RFC2681] and includes in addition any delay associated with remote endpoint processing.

Measurement of the one-way delay quantities requires that the clocks of A and B be synchronized, whereas the two-way delay metrics can be measured directly even when this is not the case (provided A and B have stable clocks).

A measurement is accomplished by sending a Delay Measurement (DM) query message over the channel to B which contains the following timestamp:

T1: the time the DM query message is transmitted from A.

When the message arrives at B, the following timestamp is recorded in the message:

T2: the time the DM query message is received at B.

At this point B transmits the message back to A, recording within it the following timestamp:

T3: the time the DM response message is transmitted from B.

When the message arrives back at A, the following timestamp is recorded in the message:

T4: the time the DM response message is received back at A.

(Strictly speaking, it is not necessary that the fourth timestamp, T4, actually be written in the message, but this is convenient for some implementations and useful if the message is to be forwarded on to an external measurement system.)

At this point, A can compute the two-way channel delay associated with the channel as

two-way channel delay = (T4 - T1) - (T3 - T2)

and the round-trip delay as

round-trip delay = T4 - T1.

If the clocks of A and B are known at A to be synchronized, then both one-way delay values, as well as the two-way channel delay, can be computed at A as

forward one-way delay = T2 - T1

reverse one-way delay = T4 - T3

two-way channel delay = forward delay + reverse delay.

Note that this formula for the two-way channel delay reduces to the one previously given, and clock synchronization is not required to compute this metric.

2.5. Delay Variation Measurement

Inter-Packet Delay Variation (IPDV) and Packet Delay Variation (PDV) [RFC5481] are performance metrics derived from one-way delay measurement and are important in some applications. IPDV represents the difference between the one-way delays of successive packets in a stream. PDV, given a measurement test interval, represents the difference between the one-way delay of a packet in the interval and that of the packet in the interval with the minimum delay.

IPDV and PDV measurements can therefore be derived from delay measurements obtained through the procedures in Section 2.4. An important point regarding delay variation measurement, however, is that it can be carried out based on one-way delay measurements even when the clocks of the two systems involved in those measurements are not synchronized with one another.

2.6. Unidirectional Measurement

In the case that the channel from A to (B1, ..., Bk) (where B2, ..., Bk refer to the point-to-multipoint case) is unidirectional, i.e. is a unidirectional LSP, LM and DM measurements can be carried out at B1, ..., Bk instead of at A.

For LM this is accomplished by initiating an LM operation at A and carrying out the same procedures as for bidirectional channels, except that no responses from B1, ..., Bk to A are generated. Instead, each terminal node B uses the A_TxP and B_RxP values in the LM messages it receives to compute the receive loss associated with the channel in essentially the same way as described previously, i.e.

B_RxLoss[n-1,n] = (A_TxP[n] - A_TxP[n-1]) - (B_RxP[n] - B_RxP[n-1])

For DM, of course, only the forward one-way delay can be measured and

the clock synchronization requirement applies.

Alternatively, if an out-of-band channel from a terminal node B back to A is available, the LM and DM message responses can be communicated to A via this channel so that the measurements can be carried out at A.

2.7. Dyadic Measurement

The basic procedures for bidirectional measurement assume that the measurement process is conducted by and for the querier node A. It is possible instead, with only minor variation of these procedures, to conduct a dyadic or "dual-ended" measurement process in which both nodes A and B perform loss or delay measurement based on the same message flow. This is achieved by stipulating that A copy the third and fourth counter or timestamp values from a response message into the third and fourth slots of the next query, which are otherwise unused, thereby providing B with equivalent information to that learned by A.

The dyadic procedure has the advantage of halving the number of messages required for both A and B to perform a given kind of measurement, but comes at the expense of each node's ability to control its own measurement process independently, and introduces additional operational complexity into the measurement protocols. The quantity of measurement traffic is also expected to be low relative to that of user traffic, particularly when 64-bit counters are used for LM. Consequently this document does not specify a dyadic operational mode. It is however still possible, and may be useful, for A to perform the extra copy, thereby providing additional information to B even when its participation in the measurement process is passive.

2.8. Loopback Measurement

Some bidirectional channels may be placed into a loopback state such that messages are looped back to the sender without modification. In this situation, LM and DM procedures can be used to carry out measurements associated with the circular path. This is done by generating "queries" with the Response flag set to 1.

For LM, the loss computation in this case is:

A_Loss[n-1,n] = (A_TxP[n] - A_TxP[n-1]) - (A_RxP[n] - A_RxP[n-1])

For DM, the round-trip delay is computed. In this case, however, the remote endpoint processing time component reflects only the time required to loop the message from channel input to channel output.

2.9. Measurement Considerations

A number of additional considerations apply in practice to the measurement methods summarized above.

2.9.1. Types of Channels

There are several types of channels in MPLS networks over which loss and delay measurement may be conducted. The channel type may restrict the kinds of measurement that can be performed. In all cases, LM and DM messages flow over the MPLS Generic Associated Channel (G-ACh), which is described in detail in [RFC5586].

Broadly, a channel in an MPLS network may be either a link, a Label Switched Path (LSP) [RFC3031], or a pseudowire [RFC3985]. Links are bidirectional and are also referred to as MPLS sections; see [RFC5586] and [RFC5960]. Pseudowires are bidirectional. Label Switched Paths may be either unidirectional or bidirectional.

The LM and DM protocols discussed in this document are initiated from a single node, the querier. A query message may be received either by a single node or by multiple nodes, depending on the nature of the channel. In the latter case these protocols provide point-to-multipoint measurement capabilities.

2.9.2. Quality of Service

Quality of Service (QoS) capabilities, in the form of the Differentiated Services architecture, apply to MPLS as specified in [RFC3270] and [RFC5462]. Different classes of traffic are distinguished by the three-bit Traffic Class (TC) field of an MPLS Label Stack Entry (LSE). Delay measurement therefore applies on a per-traffic-class basis, and the TC values of LSEs above the G-ACh Label (GAL) that precedes a DM message are significant. Packet loss can be measured with respect either to the channel as a whole or to a specific traffic class.

2.9.3. Measurement Point Location

The location of the measurement points for loss and delay within the sending and receiving nodes is implementation-dependent but directly affects the nature of the measurements. For example, a sending implementation may or may not consider a packet to be "lost", for LM purposes, that was discarded prior to transmission for queuing-related reasons; conversely, a receiving implementation may or may not consider a packet to be "lost", for LM purposes, if it was physically received but discarded during receive-path processing. The location of delay measurement points similarly determines what,

precisely, is being measured. The principal consideration here is that the behavior of an implementation in these respects MUST be made clear to the user.

2.9.4. Equal Cost Multipath

Equal Cost Multipath (ECMP) is the behavior of distributing packets across multiple alternate paths toward a destination. The use of ECMP in MPLS networks is described in BCP 128 [RFC4928]. The typical result of ECMP being performed on an LSP which is subject to delay measurement will be that only the delay of one of the available paths is and can be measured.

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.

2.9.5. Intermediate Nodes

In the case of an LSP, it may be desirable to measure the loss or delay to or from an intermediate node as well as between LSP endpoints. This can be done in principle by setting the Time to Live (TTL) field in the outer LSE appropriately when targeting a measurement message to an intermediate node. This procedure may fail, however, if hardware-assisted measurement is in use, because the processing of the packet by the intermediate node occurs only as the result of TTL expiry, and the handling of TTL expiry may occur at a later processing stage in the implementation than the hardware-assisted measurement function. Often the motivation for conducting measurements to intermediate nodes is an attempt to localize a problem that has been detected on the LSP. In this case, if intermediate nodes are not capable of performing hardware-assisted measurement, a less accurate - but usually sufficient - software-based measurement can be conducted instead.

2.9.6. Different Transmit and Receive Interfaces

The overview of the bidirectional measurement process presented in Section 2 is also applicable when the transmit and receive interfaces

at A or B differ from one another. Some additional considerations, however, do apply in this case:

- If different clocks are associated with transmit and receive processing, these clocks must be synchronized in order to compute the two-way delay.

- The DM protocol specified in this document requires that the timestamp formats used by the interfaces that receive a DM query and transmit a DM response agree.

- The LM protocol specified in this document supports both 32-bit and 64-bit counter sizes, but the use of 32-bit counters at any of the up to four interfaces involved in an LM operation will result in 32-bit LM calculations for both directions of the channel.

2.9.7. External Post-Processing

In some circumstances it may be desirable to carry out the final measurement computation at an external post-processing device dedicated to the purpose. This can be achieved in supporting implementations by, for example, configuring the querier, in the case of a bidirectional measurement session, to forward each response it receives to the post-processor via any convenient protocol. The unidirectional case can be handled similarly through configuration of the receiver, or by including an instruction in query messages for the receiver to respond out-of-band to the appropriate return address.

Post-processing devices may have the ability to store measurement data for an extended period and to generate a variety of useful statistics from them. External post-processing also allows the measurement process to be completely stateless at the querier and responder.

2.9.8. Loss Measurement Modes

The summary of loss measurement at the beginning of Section 2 above made reference to the "count of packets" transmitted and received over a channel. If the counted packets are the packets flowing over the channel in the data plane, the loss measurement is said to operate in "direct mode". If, on the other hand, the counted packets are selected control packets from which the approximate loss characteristics of the channel are being inferred, the loss measurement is said to operate in "inferred mode".

Direct LM has the advantage of being able to provide perfect loss accounting when it is available. There are, however, several

constraints associated with direct LM.

For accurate direct LM to occur, packets must not be sent between the time the transmit count for an outbound LM message is determined and the time the message is actually transmitted. Similarly, packets must not be received and processed between the time an LM message is received and the time the receive count for the message is determined. If these "synchronization conditions" do not hold, the LM message counters will not reflect the true state of the data plane, with the result that, for example, the receive count of B may be greater than the transmit count of A, and attempts to compute loss by taking the difference will yield an invalid result. This requirement for synchronization between LM message counters and the data plane may require special support from hardware-based forwarding implementations.

A limitation of direct LM is that it may be difficult or impossible to apply in cases where the channel is an LSP and the LSP label at the receiver is either nonexistent or fails to identify a unique sending node. The first case happens when Penultimate Hop Popping (PHP) is used on the LSP, and the second case generally holds for LSPs based on the Label Distribution Protocol (LDP) [RFC5036] as opposed to, for example, those based on Traffic Engineering extensions to the Resource Reservation Protocol (RSVP-TE) [RFC3209]. These conditions may make it infeasible for the receiver to identify the data-plane packets associated with a particular source and LSP in order to count them, or to infer the source and LSP context associated with an LM message. Direct LM is also vulnerable to disruption in the event that the ingress or egress interface associated with an LSP changes during the LSP's lifetime.

Inferred LM works in the same manner as direct LM except that the counted packets are special control packets, called test messages, generated by the sender. Test messages may be either packets explicitly constructed and used for LM or packets with a different primary purpose, such as those associated with a Bidirectional Forwarding Detection (BFD) [RFC5884] session.

The synchronization conditions discussed above for direct LM also apply to inferred LM, the only difference being that the required synchronization is now between the LM counters and the test message generation process. Protocol and application designers MUST take these synchronization requirements into account when developing tools for inferred LM, and make their behavior in this regard clear to the user.

Inferred LM provides only an approximate view of the loss level associated with a channel, but is typically applicable even in cases

where direct LM is not.

2.9.9. Loss Measurement Scope

In the case of direct LM, where data-plane packets are counted, there are different possibilities for which kinds of packets are included in the count and which are excluded. The set of packets counted for LM is called the loss measurement scope. As noted above, one factor affecting the LM scope is whether all data packets are counted or only those belonging to a particular traffic class. Another is whether various "auxiliary" flows associated with a data channel are counted, such as packets flowing over the G-ACh. Implementations MUST make their supported LM scopes clear to the user, and care must be taken to ensure that the scopes of the channel endpoints agree.

2.9.10. Delay Measurement Accuracy

The delay measurement procedures described in this document are designed to facilitate hardware-assisted measurement and to function in the same way whether or not such hardware assistance is used. The measurement accuracy will be determined by how closely the transmit and receive timestamps correspond to actual packet departure and arrival times.

As noted in Section 2.4, measurement of one-way delay requires clock synchronization between the devices involved, while two-way delay measurement does not involve direct comparison between non-local timestamps and thus has no synchronization requirement. The measurement accuracy will be limited by the quality of the local clock and, in the case of one-way delay measurement, by the quality of the synchonization.

2.9.11. Delay Measurement Timestamp Format

There are two significant timestamp formats in common use: the timestamp format of the Network Time Protocol (NTP), described in [RFC5905], and the timestamp format used in the IEEE 1588 Precision Time Protocol (PTP) [IEEE1588].

The NTP format has the advantages of wide use and long deployment in the Internet, and was specifically designed to make the computation of timestamp differences as simple and efficient as possible. On the other hand, there is also now a significant deployment of equipment designed to support the PTP format.

The approach taken in this document is therefore to include in DM messages fields which identify the timestamp formats used by the two devices involved in a DM operation. This implies that a node

attempting to carry out a DM operation may be faced with the problem of computing with and possibly reconciling different timestamp formats. To ensure interoperability it is necessary that support of at least one timestamp format is mandatory. This specification requires the support of the IEEE 1588 PTP format. Timestamp format support requirements are discussed in detail in Section 3.4.

3. Message Formats

Loss Measurement and Delay Measurement messages flow over the MPLS Generic Associated Channel (G-ACh) [RFC5586]. Thus, a packet containing an LM or DM message contains an MPLS label stack, with the G-ACh Label (GAL) at the bottom of the stack. The GAL is followed by an Associated Channel Header (ACH) which identifies the message type, and the message body follows the ACH.

This document defines the following ACH Channel Types:

MPLS Direct Packet Loss Measurement (DLM)
MPLS Inferred Packet Loss Measurement (ILM)
MPLS Packet Delay Measurement (DM)
MPLS Direct Packet Loss and Delay Measurement (DLM+DM)
MPLS Inferred Packet Loss and Delay Measurement (ILM+DM)

The message formats for direct and inferred LM are identical. The formats of the DLM+DM and ILM+DM messages are also identical.

For these channel types, the ACH SHALL NOT be followed by the ACH TLV Header defined in [RFC5586].

The fixed-format portion of a message MAY be followed by a block of Type-Length-Value (TLV) fields. The TLV block provides an extensible way of attaching subsidiary information to LM and DM messages. Several such TLV fields are defined below.

All integer values for fields defined in this document SHALL be encoded in network byte order.

3.1. Loss Measurement Message Format

The format of a Loss Measurement message, which follows the Associated Channel Header (ACH), is as follows:

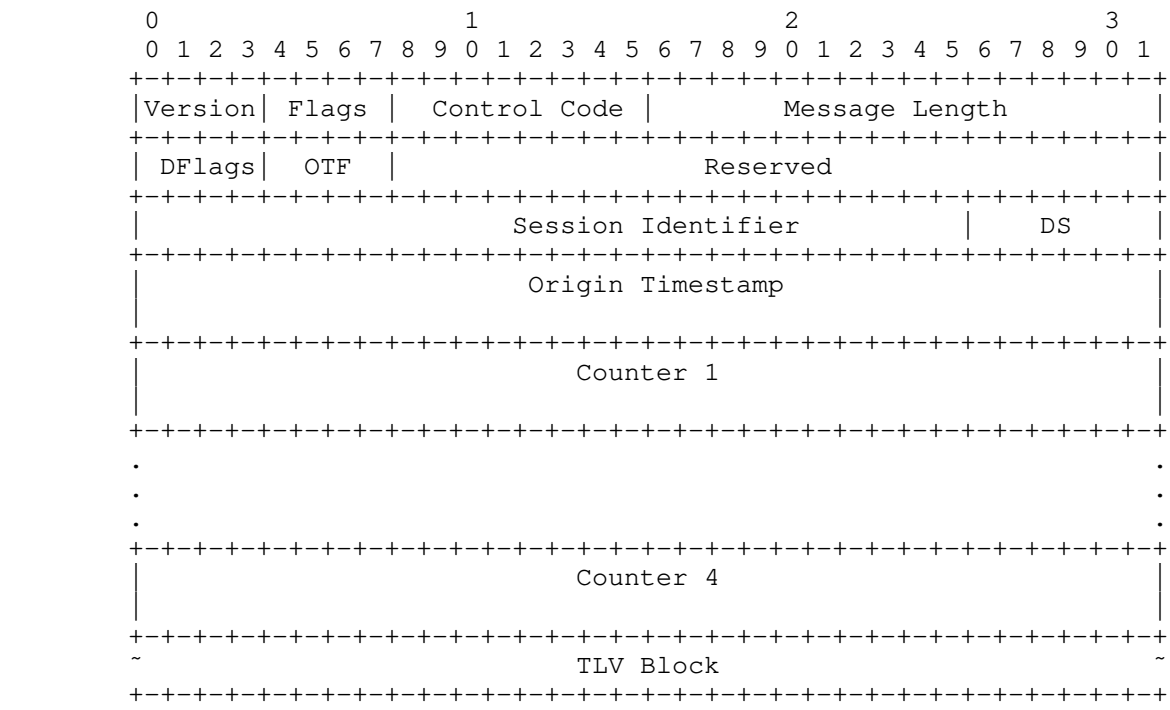


Figure 2: Loss Measurement Message Format

Reserved fields MUST be set to 0 and ignored upon receipt. The possible values for the remaining fields are as follows.

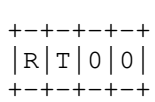
Field	Meaning
Version	Protocol version
Flags	Message control flags
Control Code	Code identifying the query or response type
Message Length	Total length of this message in bytes
Data Format Flags (DFlags)	Flags specifying the format of message data
Origin Timestamp Format (OTF)	Format of the Origin Timestamp field
Reserved	Reserved for future specification
Session Identifier	Set arbitrarily by the querier
Differentiated Services (DS) Field	Differentiated Services Code Point (DSCP) being measured
Origin Timestamp	64-bit field for query message transmission timestamp
Counter 1-4	64-bit fields for LM counter values

TLV Block Optional block of Type-Length-Value fields

The possible values for these fields are as follows.

Version: Currently set to 0.

Flags: The format of the Flags field is shown below.



Loss Measurement Message Flags

The meanings of the flag bits are:

R: Query/Response indicator. Set to 0 for a Query and 1 for a Response.

T: Traffic-class-specific measurement indicator. Set to 1 when the measurement operation is scoped to packets of a particular traffic class (DSCP value), and 0 otherwise. When set to 1, the DS field of the message indicates the measured traffic class.

0: Set to 0.

Control Code: Set as follows according to whether the message is a Query or a Response as identified by the R flag.

For a Query:

0x0: In-band Response Requested. Indicates that this query has been sent over a bidirectional channel and the response is expected over the same channel.

0x1: Out-of-band Response Requested. Indicates that the response should be sent via an out-of-band channel.

0x2: No Response Requested. Indicates that no response to the query should be sent. This mode can be used, for example, if all nodes involved are being controlled by a Network Management System.

For a Response:

Codes 0x0-0xF are reserved for non-error responses. Error response codes imply that the response does not contain valid measurement data.

0x1: Success. Indicates that the operation was successful.

0x2: Notification – Data Format Invalid. Indicates that the query was processed but the format of the data fields in this response may be inconsistent. Consequently these data fields MUST NOT be used for measurement.

0x3: Notification – Initialization In Progress. Indicates that the query was processed but this response does not contain valid measurement data because the responder's initialization process has not completed.

0x4: Notification – Data Reset Occurred. Indicates that the query was processed but a reset has recently occurred which may render the data in this response inconsistent relative to earlier responses.

0x5: Notification – Resource Temporarily Unavailable. Indicates that the query was processed but resources were unavailable to complete the requested measurement, and that consequently this response does not contain valid measurement data.

0x10: Error – Unspecified Error. Indicates that the operation failed for an unspecified reason.

0x11: Error – Unsupported Version. Indicates that the operation failed because the protocol version supplied in the query message is not supported.

0x12: Error – Unsupported Control Code. Indicates that the operation failed because the Control Code requested an operation that is not available for this channel.

0x13: Error – Unsupported Data Format. Indicates that the operation failed because the data format specified in the query is not supported.

0x14: Error – Authentication Failure. Indicates that the operation failed because the authentication data supplied in the query was missing or incorrect.

0x15: Error – Invalid Destination Node Identifier. Indicates that the operation failed because the Destination Node Identifier supplied in the query is not an identifier of this node.

0x16: Error - Connection Mismatch. Indicates that the operation failed because the channel identifier supplied in the query did not match the channel over which the query was received.

0x17: Error - Unsupported Mandatory TLV Object. Indicates that the operation failed because a TLV Object received in the query and marked as mandatory is not supported.

0x18: Error - Unsupported Query Interval. Indicates that the operation failed because the query message rate exceeded the configured threshold.

0x19: Error - Administrative Block. Indicates that the operation failed because it has been administratively disallowed.

0x1A: Error - Resource Unavailable. Indicates that the operation failed because node resources were not available.

0x1B: Error - Resource Released. Indicates that the operation failed because node resources for this measurement session were administratively released.

0x1C: Error - Invalid Message. Indicates that the operation failed because the received query message was malformed.

0x1D: Error - Protocol Error. Indicates that the operation failed because a protocol error was found in the received query message.

Message Length: Set to the total length of this message in bytes, including the Version, Flags, Control Code, and Message Length fields.

DFlags: The format of the DFlags field is shown below.

Loss Measurement Message Flags

The meanings of the DFlags bits are:

X: Extended counter format indicator. Indicates the use of extended (64-bit) counter values. Initialized to 1 upon creation (and prior to transmission) of an LM Query and copied from an LM

Query to an LM response. Set to 0 when the LM message is transmitted or received over an interface that writes 32-bit counter values.

B: Octet (byte) count. When set to 1, indicates that the Counter 1-4 fields represent octet counts. The octet count applies to all packets within the LM scope (Section 2.9.9), and the octet count of a packet sent or received over a channel includes the total length of that packet (but excludes headers, labels or framing of the channel itself). When set to 0, indicates that the Counter 1-4 fields represent packet counts.

0: Set to 0.

Origin Timestamp Format: The format of the Origin Timestamp field, as specified in Section 3.4.

Session Identifier: Set arbitrarily in a query and copied in the response, if any. This field uniquely identifies a measurement operation (also called a session) that consists of a sequence of messages. All messages in the sequence have the same Session Identifier.

DS: When the T flag is set to 1, this field is set to the DSCP value [RFC3260] that corresponds to the traffic class being measured. For MPLS, where the traffic class of a channel is identified by the three-bit Traffic Class in the channel's LSE [RFC5462], this field SHOULD be set to the Class Selector Codepoint [RFC2474] that corresponds to that Traffic Class. When the T flag is set to 0, the value of this field is arbitrary, and the field can be considered part of the Session Identifier.

Origin Timestamp: Timestamp recording the transmit time of the query message.

Counter 1-4: Referring to Section 2.2, when a query is sent from A, Counter 1 is set to A_TxP and the other counter fields are set to 0. When the query is received at B, Counter 2 is set to B_RxP. At this point, B copies Counter 1 to Counter 3 and Counter 2 to Counter 4, and re-initializes Counter 1 and Counter 2 to 0. When B transmits the response, Counter 1 is set to B_TxP. When the response is received at A, Counter 2 is set to A_RxP.

The mapping of counter types such as A_TxP to the counter fields 1-4 is designed to ensure that transmit counter values are always written at the same fixed offset in the packet, and likewise for receive counters. This property may be important for hardware processing.

When a 32-bit counter value is written to one of the counter fields, that value SHALL be written to the low-order 32 bits of the field; the high-order 32 bits of the field MUST, in this case, be set to 0.

TLV Block: Zero or more TLV fields.

3.2. Delay Measurement Message Format

The format of a Delay Measurement message, which follows the Associated Channel Header (ACH), is as follows:

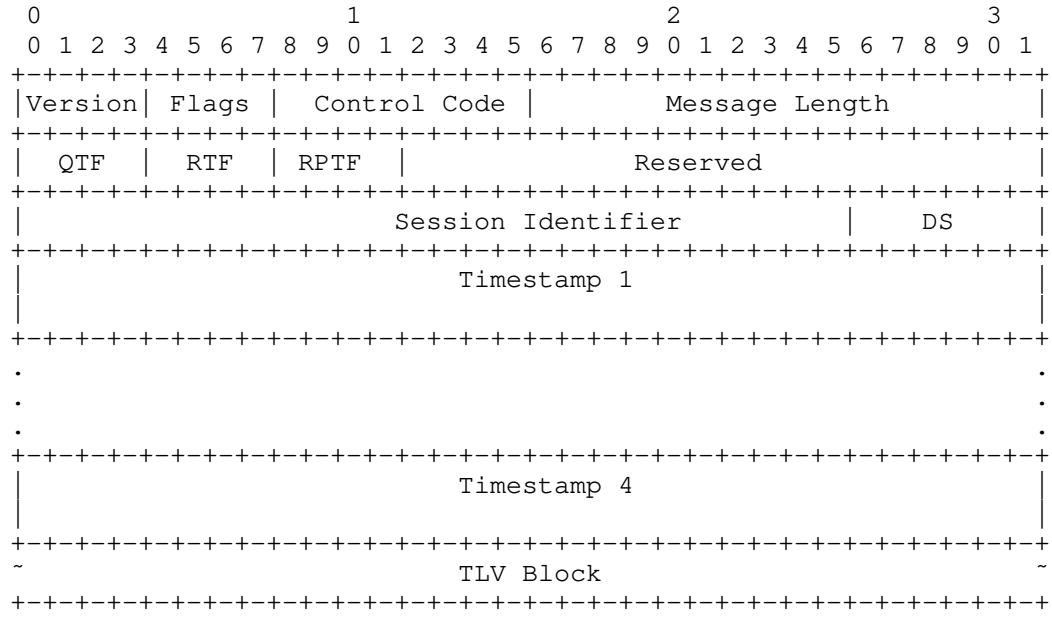


Figure 3: Delay Measurement Message Format

The meanings of the fields are summarized in the following table.

Field	Meaning
Version	Protocol version
Flags	Message control flags
Control Code	Code identifying the query or response type
Message Length	Total length of this message in bytes
QTF	Querier timestamp format
RTF	Responder timestamp format
RPTF	Responder's preferred timestamp format
Reserved	Reserved for future specification
Session Identifier	Set arbitrarily by the querier
Differentiated Services (DS) Field	Differentiated Services Code Point (DSCP) being measured
Timestamp 1-4	64-bit timestamp values
TLV Block	Optional block of Type-Length-Value fields

Reserved fields MUST be set to 0 and ignored upon receipt. The possible values for the remaining fields are as follows.

Version: Currently set to 0.

Flags: As specified in Section 3.1. The T flag in a DM message is set to 1.

Control Code: As specified in Section 3.1.

Message Length: Set to the total length of this message in bytes, including the Version, Flags, Control Code, and Message Length fields.

Querier Timestamp Format: The format of the timestamp values written by the querier, as specified in Section 3.4.

Responder Timestamp Format: The format of the timestamp values written by the responder, as specified in Section 3.4.

Responder's Preferred Timestamp Format: The timestamp format preferred by the responder, as specified in Section 3.4.

Session Identifier: As specified in Section 3.1.

DS: As specified in Section 3.1.

Timestamp 1-4: Referring to Section 2.4, when a query is sent from A, Timestamp 1 is set to T1 and the other timestamp fields are set to 0.

When the query is received at B, Timestamp 2 is set to T2. At this point, B copies Timestamp 1 to Timestamp 3 and Timestamp 2 to Timestamp 4, and re-initializes Timestamp 1 and Timestamp 2 to 0. When B transmits the response, Timestamp 1 is set to T3. When the response is received at A, Timestamp 2 is set to T4. The actual formats of the timestamp fields written by A and B are indicated by the Querier Timestamp Format and Responder Timestamp Format fields respectively.

The mapping of timestamps to the timestamp fields 1-4 is designed to ensure that transmit timestamps are always written at the same fixed offset in the packet, and likewise for receive timestamps. This property is important for hardware processing.

TLV Block: Zero or more TLV fields.

3.3. Combined Loss/Delay Measurement Message Format

The format of a combined Loss and Delay Measurement message, which follows the Associated Channel Header (ACH), is as follows:

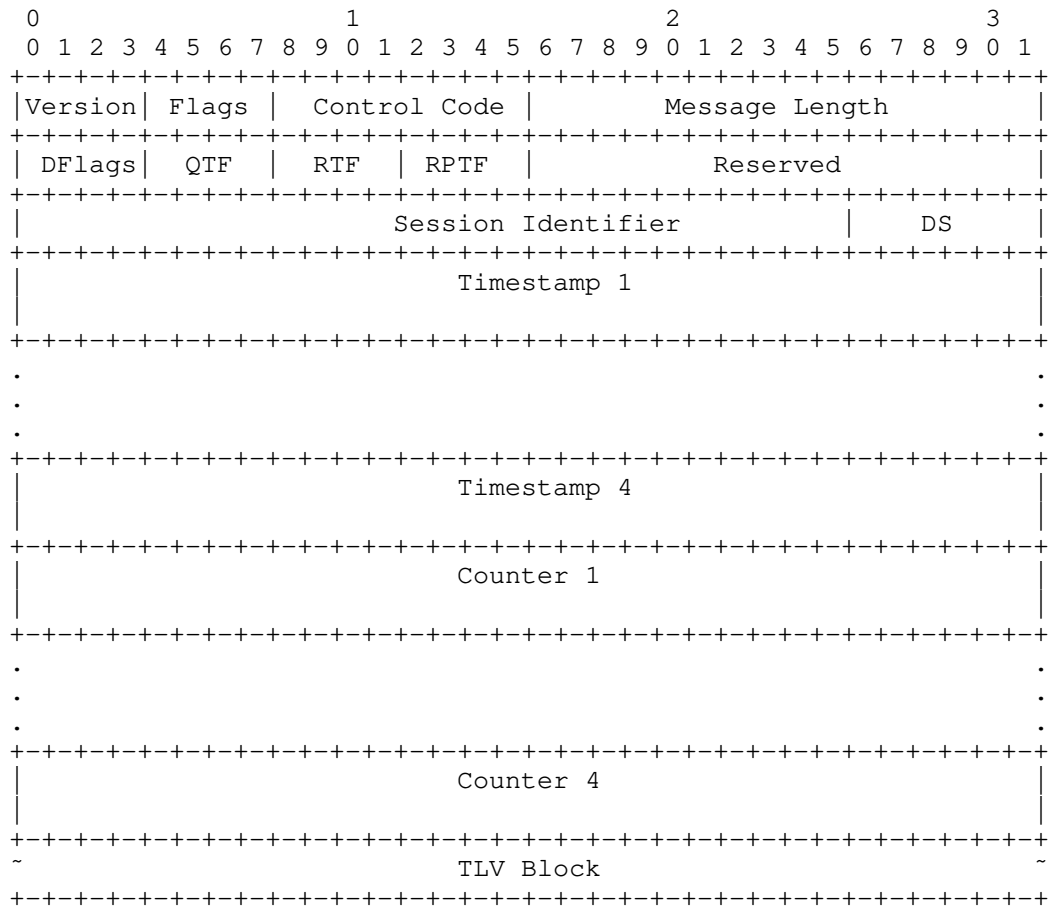


Figure 4: Loss/Delay Measurement Message Format

The fields of this message have the same meanings as the corresponding fields in the LM and DM message formats, except that the roles of the OTF and Origin Timestamp fields for LM are here played by the QTF and Timestamp 1 fields, respectively.

3.4. Timestamp Field Formats

The following timestamp format field values are specified in this document:

0: Null timestamp format. This value is a placeholder indicating that the timestamp field does not contain a meaningful timestamp.

1: Sequence number. This value indicates that the timestamp field is to be viewed as a simple 64-bit sequence number. This provides a simple solution for applications that do not require a real absolute timestamp, but only an indication of message ordering; an example is LM exception detection.

2: Network Time Protocol version 4 64-bit timestamp format [RFC5905]. This format consists of a 32-bit seconds field followed by a 32-bit fractional seconds field, so that it can be regarded as a fixed-point 64-bit quantity.

3: Low-order 64 bits of the IEEE 1588-2008 (1588v2) Precision Time Protocol timestamp format [IEEE1588]. This truncated format consists of a 32-bit seconds field followed by a 32-bit nanoseconds field, and is the same as the IEEE 1588v1 timestamp format.

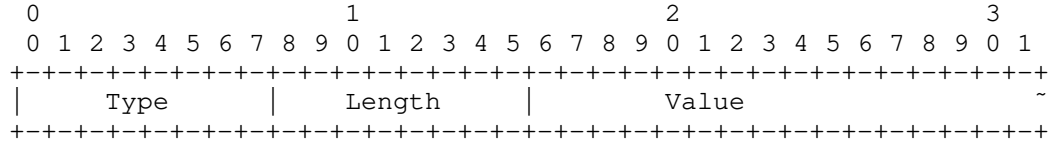
Timestamp formats of n < 64 bits in size SHALL be encoded in the 64-bit timestamp fields specified in this document using the n high-order bits of the field. The remaining 64 - n low-order bits in the field SHOULD be set to 0 and MUST be ignored when reading the field.

To ensure that it is possible to find an interoperable mode between implementations it is necessary to select one timestamp format as the default. The timestamp format chosen as the default is the truncated IEEE 1588 PTP format (format code 3 in the list above); this format MUST be supported. The rationale for this choice is discussed in Appendix A. Implementations SHOULD also be capable of reading timestamps written in NTPv4 64-bit format and reconciling them internally with PTP timestamps for measurement purposes. Support for other timestamp formats is OPTIONAL.

The implementation MUST make clear which timestamp formats it supports and the extent of its support for computation with and reconciliation of different formats for measurement purposes.

3.5. TLV Objects

The TLV Block in LM and DM messages consists of zero or more objects with the following format:



TLV Format

The Type and Length fields are each 8 bits long, and the Length field indicates the size in bytes of the Value field, which can therefore be up to 255 bytes long.

The Type space is divided into Mandatory and Optional subspaces:

Type Range	Semantics
0-127	Mandatory
128-255	Optional

Upon receipt of a query message including an unrecognized mandatory TLV object, the recipient MUST respond with an Unsupported Mandatory TLV Object error code.

The types defined are as follows:

Type	Definition
Mandatory	
0	Padding - copy in response
1	Return Address
2	Session Query Interval
3	Loopback Request
4-126	Unallocated
127	Experimental use
Optional	
128	Padding - do not copy in response
129	Destination Address
130	Source Address
131-254	Unallocated
255	Experimental use

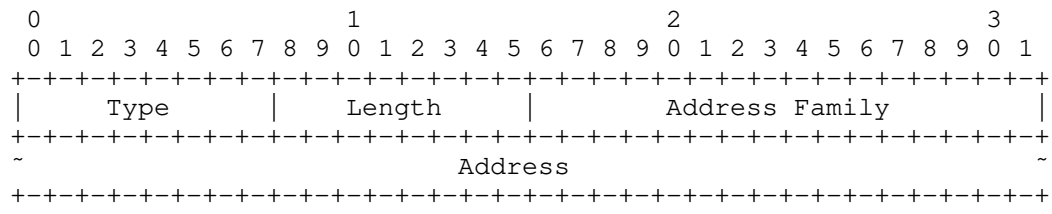
3.5.1. Padding

The two padding objects permit the augmentation of packet size; this is mainly useful for delay measurement. The type of padding indicates whether the padding supplied by the querier is to be copied to, or omitted from, the response. Asymmetrical padding may be useful when responses are delivered out-of-band or when different maximum transmission unit sizes apply to the two components of a bidirectional channel.

More than one padding object MAY be present, in which case they MUST be contiguous. The Value field of a padding object is arbitrary.

3.5.2. Addressing

The addressing objects have the following format:



Addressing Object Format

The Address Family field indicates the type of the address, and SHALL be set to one of the assigned values in the IANA Address Family Numbers registry.

The Source and Destination address objects indicate the addresses of the sender and the intended recipient of the message, respectively. The Source Address of a query message SHOULD be used as the destination for an out-of-band response unless some other out-of-band response mechanism has been configured, and unless a Return Address object is present, in which case the Return Address specifies the target of the response. The Return Address object MUST NOT appear in a response.

3.5.3. Loopback Request

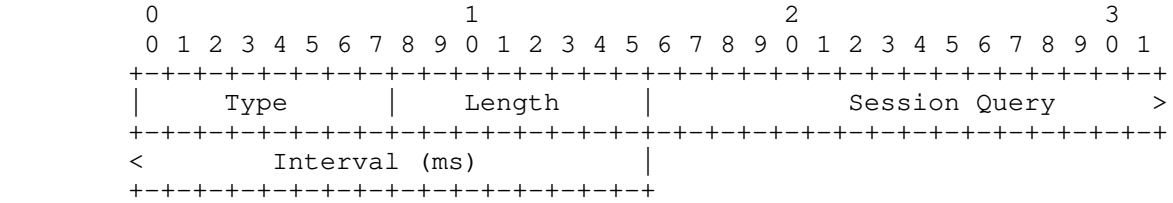
The Loopback Request object, when included in a query, indicates a request that the query message be returned to the sender unmodified. This object has a Length of 0.

Upon receiving the reflected query message back from the responder, the querier MUST NOT retransmit the message. Information that uniquely identifies the original query source, such as a Source Address object, can be included to enable the querier to differentiate one of its own loopback queries from a loopback query initiated by the far end.

This object may be useful, for example, when the querier is interested only in the round-trip delay metric. In this case no support for delay measurement is required at the responder at all, other than the ability to recognize a DM query that includes this object and return it unmodified.

3.5.4. Session Query Interval

The Value field of the Session Query Interval object is a 32-bit unsigned integer that specifies a time interval in milliseconds:



Session Query Interval Object Format

This time interval indicates the interval between successive query messages in a specific measurement session. The purpose of the Session Query Interval (SQI) object is to enable the querier and responder of a measurement session to agree on a query rate. The procedures for handling this object SHALL be as follows:

1. The querier notifies the responder that it wishes to be informed of the responder's minimum query interval for this session by including the SQI object in its query messages, with a Value of 0.

2. When the responder receives a query that includes an SQI object with a Value of 0, the responder includes an SQI object in the response with the Value set to the minimum query interval it supports for this session.

3. When the querier receives a response that includes an SQI object, it selects a query interval for the session that is greater than or equal to the Value specified in the SQI object and adjusts its query transmission rate accordingly, including in each subsequent query an SQI object with a Value equal to the selected query interval. Once a response to one of these subsequent queries has been received, the querier infers that the responder has been apprised of the selected query interval and MAY then stop including the SQI object in queries associated with this session.

Similar procedures allow the query rate to be changed during the course of the session by either the querier or the responder. For example, to inform the querier of a change in the minimum supported query interval, the responder begins including a corresponding SQI object in its responses, and the querier adjusts its query rate if necessary and includes a corresponding SQI object in its queries until a response is received.

Shorter query intervals (i.e. higher query rates) provide finer measurement granularity at the expense of additional load on measurement endpoints and the network; see Section 6 for further discussion.

4. Operation

4.1. Operational Overview

A loss or delay measurement operation, also called a session, is controlled by the querier and consists of a sequence of query messages associated with a particular channel and a common set of measurement parameters. If the session parameters include a response request, then the receiving node or nodes will (under normal conditions) generate a response message for each query message received, and these responses are also considered part of the session. All query and response messages in a session carry a common session identifier.

Measurement sessions are initiated at the discretion of the network operator and are terminated either at the operator's request or as the result of an error condition. A session may be as brief as a single message exchange, for example when a DM query is used by the operator to "ping" a remote node, or may extend throughout the lifetime of the channel.

When a session is initiated for which responses are requested, the querier SHOULD initialize a timer, called the SessionResponseTimeout, that indicates how long the querier will wait for a response before abandoning the session and notifying the user that a timeout has occurred. This timer persists for the lifetime of the session and is reset each time a response message for the session is received.

When a query message is received that requests a response, a variety of exceptional conditions may arise that prevent the responder from generating a response that contains valid measurement data. Such conditions fall broadly into two classes: transient exceptions from which recovery is possible, and fatal exceptions that require termination of the session. When an exception arises, the responder SHOULD generate a response with an appropriate Notification or Error control code according as the exception is, respectively, transient or fatal. When the querier receives an Error response, the session MUST be terminated and the user informed.

A common example of a transient exception occurs when a new session is initiated and the responder requires a period of time to become ready before it can begin providing useful responses. The response

control code corresponding to this situation is Notification - Initialization In Progress. Typical examples of fatal exceptions are cases where the querier has requested a type of measurement that the responder does not support, or where a query message is malformed.

When initiating a session the querier SHOULD employ the Session Query Interval mechanism (Section 3.5.4) to establish a mutually agreeable query rate with the responder. Responders SHOULD employ rate-limiting mechanisms to guard against the possibility of receiving an excessive quantity of query messages.

4.2. Loss Measurement Procedures

4.2.1. Initiating a Loss Measurement Operation

An LM operation for a particular channel consists of sending a sequence (LM[1], LM[2], ...) of LM query messages over the channel at a specific rate and processing the responses received, if any. As described in Section 2.2, the packet loss associated with the channel during the operation is computed as a delta between successive messages; these deltas can be accumulated to obtain a running total of the packet loss for the channel, or used to derive related metrics such as the average loss rate.

The query message transmission rate MUST be sufficiently high, given the LM message counter size (which can be either 32 or 64 bits) and the speed and minimum packet size of the underlying channel, that the ambiguity condition noted in Section 2.2 cannot arise. The implementation SHOULD assume, in evaluating this rate, that the counter size is 32 bits unless explicitly configured otherwise, or unless (in the case of a bidirectional channel) all local and remote interfaces involved in the LM operation are known to be 64-bit-capable, which can be inferred from the value of the X flag in an LM response.

4.2.2. Transmitting a Loss Measurement Query

When transmitting an LM Query, the Version field MUST be set to 0. The R flag MUST be set to 0. The T flag SHALL be set to 1 if, and only if, the measurement is specific to a particular traffic class, in which case the DS field SHALL identify that traffic class.

The X flag MUST be set to 1 if the transmitting interface writes 64-bit LM counters, and otherwise MUST be set to 0 to indicate that 32-bit counters are written. The B flag SHALL be set to 1 to indicate that the counter fields contain octet counts, or to 0 to indicate packet counts.

The Control Code field MUST be set to one of the values for Query messages listed in Section 3.1; if the channel is unidirectional, this field MUST NOT be set to 0x0 (Query: in-band response requested).

The Session Identifier field can be set arbitrarily.

The Origin Timestamp field SHALL be set to the time at which this message is transmitted, and the Origin Timestamp Format field MUST be set to indicate its format, according to Section 3.4.

The Counter 1 field SHOULD be set to the total count of units (packets or octets, according to the B flag) transmitted over the channel prior to this LM Query, or to 0 if this is the beginning of a measurement session for which counter data is not yet available. The Counter 2 field MUST be set to 0. If a response was previously received in this measurement session, the Counter 1 and Counter 2 fields of the most recent such response MAY be copied to the Counter 3 and Counter 4 fields, respectively, of this query; otherwise, the Counter 3 and Counter 4 fields MUST be set to 0.

4.2.3. Receiving a Loss Measurement Query

Upon receipt of an LM Query message, the Counter 2 field SHOULD be set to the total count of units (packets or octets, according to the B flag) received over the channel prior to this LM Query. If the receiving interface writes 32-bit LM counters, the X flag MUST be set to 0.

At this point the LM Query message must be inspected. If the Control Code field is set to 0x2 (no response requested), an LM Response message MUST NOT be transmitted. If the Control Code field is set to 0x0 (in-band response requested) or 0x1 (out-of-band response requested), then an in-band or out-of-band response, respectively, SHOULD be transmitted unless this has been prevented by an administrative, security or congestion control mechanism.

In the case of a fatal exception that prevents the requested measurement from being made, the error SHOULD be reported, either via a response if one was requested or else as a notification to the user.

4.2.4. Transmitting a Loss Measurement Response

When constructing a Response to an LM Query, the Version field MUST be set to 0. The R flag MUST be set to 1. The value of the T flag MUST be copied from the LM Query.

The X flag MUST be set to 0 if the transmitting interface writes 32-bit LM counters; otherwise its value MUST be copied from the LM Query. The B flag MUST be copied from the LM Query.

The Session Identifier, Origin Timestamp, and Origin Timestamp Format fields MUST be copied from the LM Query. The Counter 1 and Counter 2 fields from the LM Query MUST be copied to the Counter 3 and Counter 4 fields, respectively, of the LM Response.

The Control Code field MUST be set to one of the values for Response messages listed in Section 3.1. The value 0x10 (Unspecified Error) SHOULD NOT be used if one of the other more specific error codes is applicable.

If the response is transmitted in-band, the Counter 1 field SHOULD be set to the total count of units transmitted over the channel prior to this LM Response. If the response is transmitted out-of-band, the Counter 1 field MUST be set to 0. In either case, the Counter 2 field MUST be set to 0.

4.2.5. Receiving a Loss Measurement Response

Upon in-band receipt of an LM Response message, the Counter 2 field is set to the total count of units received over the channel prior to this LM Response. If the receiving interface writes 32-bit LM counters, the X flag is set to 0. (Since the life of the LM message in the network has ended at this point, it is up to the receiver whether these final modifications are made to the packet. If the message is to be forwarded on for external post-processing (Section 2.9.7) then these modifications MUST be made.)

Upon out-of-band receipt of an LM Response message, the Counter 1 and Counter 2 fields MUST NOT be used for purposes of loss measurement.

If the Control Code in an LM Response is anything other than 0x1 (Success), the counter values in the response MUST NOT be used for purposes of loss measurement. If the Control Code indicates an error condition, or if the response message is invalid, the LM operation MUST be terminated and an appropriate notification to the user generated.

4.2.6. Loss Calculation

Calculation of packet loss is carried out according to the procedures in Section 2.2. The X flag in an LM message informs the device performing the calculation whether to perform 32-bit or 64-bit arithmetic. If the flag value is equal to 1, all interfaces involved in the LM operation have written 64-bit counter values, and 64-bit

arithmetic can be used. If the flag value is equal to 0, at least one interface involved in the operation has written a 32-bit counter value, and 32-bit arithmetic is carried out using the low-order 32 bits of each counter value.

Note that the semantics of the X flag allow all devices to interoperate regardless of their counter size support. Thus, an implementation MUST NOT generate an error response based on the value of this flag.

4.2.7. Quality of Service

The TC field of the LSE corresponding to the channel (e.g. LSP) being measured SHOULD be set to a traffic class equal to or better than the best TC within the measurement scope to minimize the chance of out-of-order conditions.

4.2.8. G-ACh Packets

By default, direct LM MUST exclude packets transmitted and received over the Generic Associated Channel (G-ACh). An implementation MAY provide the means to alter the direct LM scope to include some or all G-ACh messages. Care must be taken when altering the LM scope to ensure that both endpoints are in agreement.

4.2.9. Test Messages

In the case of inferred LM, the packets counted for LM consist of test messages generated for this purpose, or of some other class of packets deemed to provide a good proxy for data packets flowing over the channel. The specification of test protocols and proxy packets is outside the scope of this document, but some guidelines are discussed below.

An identifier common to both the test or proxy messages and the LM messages may be required to make correlation possible. The combined value of the Session Identifier and DS fields SHOULD be used for this purpose when possible. That is, test messages in this case will include a 32-bit field which can carry the value of the combined Session Identifier + DS field present in LM messages. When TC-specific LM is conducted, the DS field of the LSE in the label stack of a test message corresponding to the channel (e.g. LSP) over which the message is sent MUST correspond to the DS value in the associated LM messages.

A separate test message protocol SHOULD include a timeout value in its messages that informs the responder when to discard any state associated with a specific test.

4.2.10. Message Loss and Packet Misorder Conditions

Because an LM operation consists of a message sequence with state maintained from one message to the next, LM is subject to the effects of lost messages and misordered packets in a way that DM is not. Because this state exists only on the querier, the handling of these conditions is, strictly speaking, a local matter. This section, however, presents recommended procedures for handling such conditions. Note that in the absence of ECMP, packet misordering within a traffic class is a relatively rare event.

The first kind of anomaly that may occur is that one or more LM messages may be lost in transit. The effect of such loss is that when an LM Response is next received at the querier, an unambiguous interpretation of the counter values it contains may be impossible, for the reasons described at the end of Section 2.2. Whether this is so depends on the number of messages lost and the other variables mentioned in that section, such as the LM message rate and the channel parameters.

Another possibility is that LM messages are misordered in transit, so that for instance the response to LM[n] is received prior to the response to LM[n-1]. A typical implementation will discard the late response to LM[n-1], so that the effect is the same as the case of a lost message.

Finally, LM is subject to the possibility that data packets are misordered relative to LM messages. This condition can result, for example, in a transmit count of 100 and a corresponding receive count of 101. The effect here is that the A_TxLoss[n-1,n] value (for example) for a given measurement interval will appear to be extremely (if not impossibly) large. The other case, where an LM message arrives earlier than some of the packets, simply results in those packets being counted as lost.

An implementation SHOULD identify a threshold value that indicates the upper bound of lost packets measured in a single computation beyond which the interval is considered unmeasurable. This is called the MaxLMIntervalLoss threshold. It is clear that this threshold should be no higher than the maximum number of packets (or bytes) the channel is capable of transmitting over the interval, but it may be lower. Upon encountering an unmeasurable interval, the LM state (i.e. data values from the last LM message received) SHOULD be discarded.

With regard to lost LM messages, the MaxLMInterval (see Section 2.2) indicates the maximum amount of time that can elapse before the LM state is discarded. If some messages are lost, but a message is

subsequently received within MaxLMInterval, its timestamp or sequence number will quantify the loss, and it MAY still be used for measurement, although the measurement interval will in this case be longer than usual.

If an LM message is received that has a timestamp less than or equal to the timestamp of the last LM message received, this indicates that an exception has occurred, and the current interval SHOULD be considered unmeasurable unless the implementation has some other way of handling this condition.

4.3. Delay Measurement Procedures

4.3.1. Transmitting a Delay Measurement Query

When transmitting a DM Query, the Version and Reserved fields MUST be set to 0. The R flag MUST be set to 0, the T flag MUST be set to 1, and the remaining flag bits MUST be set to 0.

The Control Code field MUST be set to one of the values for Query messages listed in Section 3.1; if the channel is unidirectional, this field MUST NOT be set to 0x0 (Query: in-band response requested).

The Querier Timestamp Format field MUST be set to the timestamp format used by the querier when writing timestamp fields in this message; the possible values for this field are listed in Section 3.4. The Responder Timestamp Format and Responder's Preferred Timestamp Format fields MUST be set to 0.

The Session Identifier field can be set arbitrarily. The DS field MUST be set to the traffic class being measured.

The Timestamp 1 field SHOULD be set to the time at which this DM Query is transmitted, in the format indicated by the Querier Timestamp Format field. The Timestamp 2 field MUST be set to 0. If a response was previously received in this measurement session, the Timestamp 1 and Timestamp 2 fields of the most recent such response MAY be copied to the Timestamp 3 and Timestamp 4 fields, respectively, of this query; otherwise, the Timestamp 3 and Timestamp 4 fields MUST be set to 0.

4.3.2. Receiving a Delay Measurement Query

Upon receipt of a DM Query message, the Timestamp 2 field SHOULD be set to the time at which this DM Query is received.

At this point the DM Query message must be inspected. If the Control

Code field is set to 0x2 (no response requested), a DM Response message MUST NOT be transmitted. If the Control Code field is set to 0x0 (in-band response requested) or 0x1 (out-of-band response requested), then an in-band or out-of-band response, respectively, SHOULD be transmitted unless this has been prevented by an administrative, security or congestion control mechanism.

In the case of a fatal exception that prevents the requested measurement from being made, the error SHOULD be reported, either via a response if one was requested or else as a notification to the user.

4.3.3. Transmitting a Delay Measurement Response

When constructing a Response to a DM Query, the Version and Reserved fields MUST be set to 0. The R flag MUST be set to 1, the T flag MUST be set to 1, and the remaining flag bits MUST be set to 0.

The Session Identifier and Querier Timestamp Format (QTF) fields MUST be copied from the DM Query. The Timestamp 1 and Timestamp 2 fields from the DM Query MUST be copied to the Timestamp 3 and Timestamp 4 fields, respectively, of the DM Response.

The Responder Timestamp Format (RTF) field MUST be set to the timestamp format used by the responder when writing timestamp fields in this message, i.e. Timestamp 4 and (if applicable) Timestamp 1; the possible values for this field are listed in Section 3.4. Furthermore, the RTF field MUST be set equal either to the QTF or the RPTF field. See Section 4.3.5 for guidelines on selection of the value for this field.

The Responder's Preferred Timestamp Format (RPTF) field MUST be set to one of the values listed in Section 3.4 and SHOULD be set to indicate the timestamp format with which the responder can provide the best accuracy for purposes of delay measurement.

The Control Code field MUST be set to one of the values for Response messages listed in Section 3.1. The value 0x10 (Unspecified Error) SHOULD NOT be used if one of the other more specific error codes is applicable.

If the response is transmitted in-band, the Timestamp 1 field SHOULD be set to the time at which this DM Response is transmitted. If the response is transmitted out-of-band, the Timestamp 1 field MUST be set to 0. In either case, the Timestamp 2 field MUST be set to 0.

If the response is transmitted in-band and the Control Code in the message is 0x1 (Success), then the Timestamp 1 and Timestamp 4 fields

MUST have the same format, which will be the format indicated in the Responder Timestamp Format field.

4.3.4. Receiving a Delay Measurement Response

Upon in-band receipt of a DM Response message, the Timestamp 2 field is set to the time at which this DM Response is received. (Since the life of the DM message in the network has ended at this point, it is up to the receiver whether this final modification is made to the packet. If the message is to be forwarded on for external post-processing (Section 2.9.7) then these modifications MUST be made.)

Upon out-of-band receipt of a DM Response message, the Timestamp 1 and Timestamp 2 fields MUST NOT be used for purposes of delay measurement.

If the Control Code in a DM Response is anything other than 0x1 (Success), the timestamp values in the response MUST NOT be used for purposes of delay measurement. If the Control Code indicates an error condition, or if the response message is invalid, the DM operation MUST be terminated and an appropriate notification to the user generated.

4.3.5. Timestamp Format Negotiation

In case either the querier or the responder in a DM transaction is capable of supporting multiple timestamp formats, it is desirable to determine the optimal format for purposes of delay measurement on a particular channel. The procedures for making this determination SHALL be as follows.

Upon sending an initial DM Query over a channel, the querier sets the Querier Timestamp Format (QTF) field to its preferred timestamp format.

Upon receiving any DM Query message, the responder determines whether it is capable of writing timestamps in the format specified by the QTF field. If so, the Responder Timestamp Format (RTF) field is set equal to the QTF field. If not, the RTF field is set equal to the Responder's Preferred Timestamp Format (RPTF) field.

The process of changing from one timestamp format to another at the responder may result in the Timestamp 1 and Timestamp 4 fields in an in-band DM Response having different formats. If this is the case, the Control Code in the response MUST NOT be set to 0x1 (Success). Unless an error condition has occurred, the Control Code MUST be set to 0x2 (Notification - Data Format Invalid).

Upon receiving a DM Response, the querier knows from the RTF field in the message whether the responder is capable of supporting its preferred timestamp format: if it is, the RTF will be equal to the QTF. The querier also knows the responder's preferred timestamp format from the RPTF field. The querier can then decide whether to retain its current QTF or to change it and repeat the negotiation procedures.

4.3.5.1. Single-Format Procedures

When an implementation supports only one timestamp format, the procedures above reduce to the following simple behavior:

- All DM Queries are transmitted with the same QTF;
- All DM Responses are transmitted with the same RTF, and the RPTF is always set equal to the RTF;
- All DM Responses received with RTF not equal to QTF are discarded;
- On a unidirectional channel, all DM Queries received with QTF not equal to the supported format are discarded.

4.3.6. Quality of Service

The TC field of the LSE corresponding to the channel (e.g. LSP) being measured MUST be set to the value that corresponds to the DS field in the DM message.

4.4. Combined Loss/Delay Measurement Procedures

The combined LM/DM message defined in Section 3.3 allows loss and delay measurement to be carried out simultaneously. This message SHOULD be treated as an LM message which happens to carry additional timestamp data, with the timestamp fields processed as per delay measurement procedures.

5. Implementation Disclosure Requirements

This section summarizes the requirements placed on implementations for capabilities disclosure. The purpose of these requirements is to ensure that end users have a clear understanding of implementation capabilities and characteristics that have a direct impact on how loss and delay measurement mechanisms function in specific situations. Implementations are REQUIRED to state:

- METRICS: Which of the following metrics are supported: packet loss, packet throughput, octet loss, octet throughput, average loss rate, one-way delay, round-trip delay, two-way channel delay, packet delay variation.

- MP-LOCATION: The location of loss and delay measurement points with respect to other stages of packet processing, such as queuing.

- CHANNEL-TYPES: The types of channels for which LM and DM are supported, including LSP types, pseudowires, and sections (links).

- QUERY-RATE: The minimum supported query intervals for LM and DM sessions, both in the querier and responder roles.

- LOOP: Whether loopback measurement (Section 2.8) is supported.

- LM-TYPES: Whether direct or inferred LM is supported, and for the latter, which test protocols or proxy message types are supported.

- LM-COUNTERS: Whether 64-bit counters are supported.

- LM-ACCURACY: The expected measurement accuracy levels for the supported forms of LM, and the expected impact of exception conditions such as lost and misordered messages.

- LM-SYNC: The implementation's behavior in regard to the synchronization conditions discussed in Section 2.9.8.

- LM-SCOPE: The supported LM scopes (Section 2.9.9 and Section 4.2.8).

- DM-ACCURACY: The expected measurement accuracy levels for the supported forms of DM.

- DM-TS-FORMATS: The supported timestamp formats and the extent of support for computation with and reconciliation of different formats.

6. Congestion Considerations

An MPLS network may be traffic-engineered in such a way that the bandwidth required both for client traffic and for control, management and OAM traffic is always available. The following congestion considerations therefore apply only when this is not the case.

The proactive generation of Loss Measurement and Delay Measurement messages for purposes of monitoring the performance of an MPLS channel naturally results in a degree of additional load placed on both the network and the terminal nodes of the channel. When configuring such monitoring, operators should be mindful of the overhead involved and should choose transmit rates that do not stress network resources unduly; such choices must be informed by the deployment context. In case of slower links or lower-speed devices, for example, lower Loss Measurement message rates can be chosen, up to the limits noted at the end of Section 2.2.

In general, lower measurement message rates place less load on the network at the expense of reduced granularity. For delay measurement this reduced granularity translates to a greater possibility that the delay associated with a channel temporarily exceeds the expected threshold without detection. For loss measurement, it translates to a larger gap in loss information in case of exceptional circumstances such as lost LM messages or misordered packets.

When carrying out a sustained measurement operation such as an LM operation or continuous pro-active DM operation, the querier SHOULD take note of the number of lost measurement messages (queries for which a response is never received) and set a corresponding Measurement Message Loss Threshold. If this threshold is exceeded, the measurement operation SHOULD be suspended so as not to exacerbate the possible congestion condition. This suspension SHOULD be accompanied by an appropriate notification to the user so that the condition can be investigated and corrected.

From the receiver perspective, the main consideration is the possibility of receiving an excessive quantity of measurement messages. An implementation SHOULD employ a mechanism such as rate-limiting to guard against the effects of this case.

7. Manageability Considerations

The measurement protocols described in this document are intended to serve as infrastructure to support a wide range of higher-level monitoring and diagnostic applications, from simple command-line diagnostic tools to comprehensive network performance monitoring and analysis packages. The specific mechanisms and considerations for protocol configuration, initialization and reporting thus depend on the nature of the application.

In the case of on-demand diagnostics, the diagnostic application may provide parameters such as the measurement type, the channel, the query rate, and the test duration when initiating the diagnostic;

results and exception conditions are then reported directly to the application. The system may discard the statistics accumulated during the test after the results have been reported, or retain them to provide a historical measurement record.

Alternatively, measurement configuration may be supplied as part of the channel configuration itself in order to support continuous monitoring of the channel's performance characteristics. In this case the configuration will typically include quality thresholds depending on the service-level agreement, the crossing of which will trigger warnings or alarms, and result reporting and exception notification will be integrated into the system-wide network management and reporting framework.

8. Security Considerations

This document describes procedures for the measurement of performance metrics over a pre-existing MPLS path (a pseudowire, LSP, or section). As such it assumes that a node involved in a measurement operation has previously verified the integrity of the path and the identity of the far end using existing MPLS mechanisms such as Bidirectional Forwarding Detection (BFD) [RFC5884]; tools, techniques, and considerations for securing MPLS paths are discussed in detail in [RFC5920].

When such mechanisms are not available, and where security of the measurement operation is a concern, reception of Generic Associated Channel messages with the Channel Types specified in this document SHOULD be disabled. Implementations MUST provide the ability to disable these protocols on a per-Channel-Type basis.

Even when the identity of the far end has been verified, the measurement protocols remain vulnerable to injection and man-in-the-middle attacks. The impact of such an attack would be to compromise the quality of performance measurements on the affected path. An attacker positioned to disrupt these measurements is, however, capable of causing much greater damage by disrupting far more critical elements of the network such as the network control plane or user traffic flows. A disruption of the measurement protocols would at worst interfere with the monitoring of the performance aspects of the service level agreement associated with the path; the existence of such a disruption would imply that a much more serious breach of basic path integrity had already occurred.

Such attacks can be mitigated if desired by performing basic validation and sanity checks, at the querier, of the counter or timestamp fields in received measurement response messages. The

minimal state associated with these protocols also limits the extent of measurement disruption that can be caused by a corrupt or invalid message to a single query/response cycle.

Cryptographic mechanisms capable of signing or encrypting the contents of the measurement packets without degrading the measurement performance are not currently available. In light of the preceding discussion, the absence of such cryptographic mechanisms does not raise significant security issues.

Users concerned with the security of out-of-band responses over IP networks SHOULD employ suitable security mechanisms such as IPsec [RFC4301] to protect the integrity of the return path.

9. IANA Considerations

This document makes the following requests of IANA:

- o Allocation of Channel Types in the PW Associated Channel Type registry
- o Creation of a Measurement Timestamp Type registry
- o Creation of an MPLS Loss/Delay Measurement Control Code registry
- o Creation of an MPLS Loss/Delay Measurement Type-Length-Value (TLV) Object registry

9.1. Allocation of PW Associated Channel Types

As per the IANA considerations in [RFC5586], IANA is requested to allocate the following Channel Types in the PW Associated Channel Type registry:

Value	Description	TLV Follows	Reference
TBD	MPLS Direct Packet Loss Measurement (DLM)	No	(this draft)
TBD	MPLS Inferred Packet Loss Measurement (ILM)	No	(this draft)
TBD	MPLS Packet Delay Measurement (DM)	No	(this draft)
TBD	MPLS Direct Packet Loss and Delay Measurement (DLM+DM)	No	(this draft)
TBD	MPLS Inferred Packet Loss and Delay Measurement (ILM+DM)	No	(this draft)

The values marked TBD are to be allocated by IANA as appropriate.

9.2. Creation of Measurement Timestamp Type Registry

IANA is requested to create a new Measurement Timestamp Type registry, with format and initial allocations as follows:

Type	Description	Size in bits	Reference
0	Null Timestamp	64	(this draft)
1	Sequence Number	64	(this draft)
2	Network Time Protocol version 4 64-bit Timestamp	64	(this draft)
3	Truncated IEEE 1588v2 PTP Timestamp	64	(this draft)

The range of the Type field is 0-15.

The allocation policy for this registry is IETF Review.

9.3. Creation of MPLS Loss/Delay Measurement Control Code Registry

IANA is requested to create a new MPLS Loss/Delay Measurement Control Code registry. This registry is divided into two separate parts, one for Query Codes and the other for Response Codes, with formats and initial allocations as follows:

Query Codes

Code	Description	Reference
0x0	In-band Response Requested	(this draft)
0x1	Out-of-band Response Requested	(this draft)
0x2	No Response Requested	(this draft)

Response Codes

Code	Description	Reference
0x0	Reserved	(this draft)
0x1	Success	(this draft)
0x2	Data Format Invalid	(this draft)
0x3	Initialization In Progress	(this draft)
0x4	Data Reset Occurred	(this draft)
0x5	Resource Temporarily Unavailable	(this draft)
0x10	Unspecified Error	(this draft)
0x11	Unsupported Version	(this draft)
0x12	Unsupported Control Code	(this draft)
0x13	Unsupported Data Format	(this draft)
0x14	Authentication Failure	(this draft)
0x15	Invalid Destination Node Identifier	(this draft)
0x16	Connection Mismatch	(this draft)
0x17	Unsupported Mandatory TLV Object	(this draft)
0x18	Unsupported Query Interval	(this draft)
0x19	Administrative Block	(this draft)
0x1A	Resource Unavailable	(this draft)
0x1B	Resource Released	(this draft)
0x1C	Invalid Message	(this draft)
0x1D	Protocol Error	(this draft)

IANA is also requested to indicate that the values 0x0 - 0xF in the Response Code section are reserved for non-error response codes.

The range of the Code field is 0 - 255.

The allocation policy for this registry is IETF Review.

9.4. Creation of MPLS Loss/Delay Measurement TLV Object Registry

IANA is requested to create a new MPLS Loss/Delay Measurement TLV Object registry, with format and initial allocations as follows:

Type	Description	Reference
0	Padding - copy in response	(this draft)
1	Return Address	(this draft)
2	Session Query Interval	(this draft)
3	Loopback Request	(this draft)
127	Experimental use	(this draft)
128	Padding - do not copy in response	(this draft)
129	Destination Address	(this draft)
130	Source Address	(this draft)
255	Experimental use	(this draft)

IANA is also requested to indicate that Types 0-127 are classified as Mandatory, and that Types 128-255 are classified as Optional.

The range of the Type field is 0 - 255.

The allocation policy for this registry is IETF Review.

10. Acknowledgments

The authors wish to thank the many participants of the MPLS working group who provided detailed review and feedback on this document. The authors offer special thanks to Alexander Vainshtein, Loa Andersson, and Hiroyuki Takagi for many helpful thoughts and discussions, to Linda Dunbar for the idea of using LM messages for throughput measurement, and to Ben Niven-Jenkins, Marc Lasserre, and Ben Mack-Crane for their valuable comments.

11. References

11.1. Normative References

[IEEE1588] IEEE, "1588-2008 IEEE Standard for a Precision Clock Synchronization Protocol for Networked Measurement and Control Systems", March 2008.

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

[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.

[RFC3031] Rosen, E., Viswanathan, A., and R. Callon, "Multiprotocol Label Switching Architecture", RFC 3031, January 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.

[RFC5462] Andersson, L. and R. Asati, "Multiprotocol Label Switching (MPLS) Label Stack Entry: "EXP" Field Renamed to "Traffic Class" Field", RFC 5462, February 2009.

[RFC5586] Bocci, M., Vigoureux, M., and S. Bryant, "MPLS Generic Associated Channel", RFC 5586, June 2009.

[RFC5905] Mills, D., Martin, J., Burbank, J., and W. Kasch, "Network Time Protocol Version 4: Protocol and Algorithms Specification", RFC 5905, June 2010.

11.2. Informative References

[I-D.ietf-mpls-tp-loss-delay-profile]
Frost, D. and S. Bryant, "A Packet Loss and Delay Measurement Profile for MPLS-based Transport Networks", draft-ietf-mpls-tp-loss-delay-profile-03 (work in progress), April 2011.

[RFC2679] Almes, G., Kalidindi, S., and M. Zekauskas, "A One-way Delay Metric for IPPM", RFC 2679, September 1999.

[RFC2680] Almes, G., Kalidindi, S., and M. Zekauskas, "A One-way Packet Loss Metric for IPPM", RFC 2680, September 1999.

[RFC2681] Almes, G., Kalidindi, S., and M. Zekauskas, "A Round-trip Delay Metric for IPPM", RFC 2681, September 1999.

[RFC3209] Awduche, D., Berger, L., Gan, D., Li, T., Srinivasan, V., and G. Swallow, "RSVP-TE: Extensions to RSVP for LSP Tunnels", RFC 3209, December 2001.

[RFC3260] Grossman, D., "New Terminology and Clarifications for Diffserv", RFC 3260, April 2002.

[RFC3985] Bryant, S. and P. Pate, "Pseudo Wire Emulation Edge-to-Edge (PWE3) Architecture", RFC 3985, March 2005.

[RFC4301] Kent, S. and K. Seo, "Security Architecture for the Internet Protocol", RFC 4301, December 2005.

[RFC4656] Shalunov, S., Teitelbaum, B., Karp, A., Boote, J., and M. Zekauskas, "A One-way Active Measurement Protocol (OWAMP)", RFC 4656, September 2006.

[RFC4928] Swallow, G., Bryant, S., and L. Andersson, "Avoiding Equal Cost Multipath Treatment in MPLS Networks", BCP 128, RFC 4928, June 2007.

[RFC5036] Andersson, L., Minei, I., and B. Thomas, "LDP Specification", RFC 5036, October 2007.

[RFC5357] Hedayat, K., Krzanowski, R., Morton, A., Yum, K., and J. Babiarz, "A Two-Way Active Measurement Protocol (TWAMP)", RFC 5357, October 2008.

[RFC5481] Morton, A. and B. Claise, "Packet Delay Variation Applicability Statement", RFC 5481, March 2009.

[RFC5884] Aggarwal, R., Kompella, K., Nadeau, T., and G. Swallow, "Bidirectional Forwarding Detection (BFD) for MPLS Label Switched Paths (LSPs)", RFC 5884, June 2010.

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

[RFC5921] Bocci, M., Bryant, S., Frost, D., Levrau, L., and L. Berger, "A Framework for MPLS in Transport Networks", RFC 5921, July 2010.

[RFC5960] Frost, D., Bryant, S., and M. Bocci, "MPLS Transport Profile Data Plane Architecture", RFC 5960, August 2010.

[Y.1731] ITU-T Recommendation Y.1731, "OAM Functions and Mechanisms for Ethernet based Networks", February 2008.

Appendix A. Default Timestamp Format Rationale

This document initially proposed the Network Time Protocol (NTP) timestamp format as the mandatory default, as this is the normal default timestamp in IETF protocols and thus would seem the "natural" choice. However a number of considerations have led instead to the specification of the truncated IEEE 1588 Precision Time Protocol (PTP) timestamp as the default. NTP has not gained traction in industry as the protocol of choice for high quality timing infrastructure, whilst IEEE 1588 PTP has become the de facto time transfer protocol in networks which are specially engineered to provide high accuracy time distribution service. The PTP timestamp

format is also the ITU-T format of choice for packet transport networks, which may rely on MPLS protocols. Applications such as one-way delay measurement need the best time service available, and converting between the NTP and PTP timestamp formats is not a trivial transformation, particularly when it is required that this be done in real time without loss of accuracy.

The truncated IEEE 1588 PTP format specified in this document is considered to provide a more than adequate wrap time and greater time resolution than it is expected will be needed for the operational lifetime of this protocol. By truncating the timestamp at both the high and low order bits, the protocol achieves a worthwhile reduction in system resources.

Authors' Addresses

Dan Frost
Cisco Systems

Email: danfrost@cisco.com

Stewart Bryant
Cisco Systems

Email: stbryant@cisco.com

MPLS Working Group
Internet-Draft
Intended status: Standards Track
Expires: May 26, 2016

E. Bellagamba
G. Mirsky
Ericsson
L. Andersson
Huawei Technologies
P. Skoldstrom
Acreo AB
D. Ward
Cisco
J. Drake
Juniper

November 23, 2015

Configuration of Proactive Operations, Administration, and Maintenance (OAM) Functions for MPLS-based Transport Networks using LSP Ping
draft-ietf-mpls-lsp-ping-mpls-tp-oam-conf-16

Abstract

This specification describes the configuration of proactive MPLS-TP Operations, Administration, and Maintenance (OAM) Functions for a given Label Switched Path (LSP) using a set of TLVs that are carried by the LSP-Ping protocol.

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 26, 2016.

Copyright Notice

Copyright (c) 2015 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

The MPLS Transport Profile (MPLS-TP) describes a profile of MPLS that enables operational models typical in transport networks, while providing additional Operations, Administration, and Maintenance (OAM), survivability and other maintenance functions not currently supported by MPLS. [RFC5860] defines the requirements for the OAM functionality of MPLS-TP.

This document describes the configuration of proactive MPLS-TP OAM Functions for a given Label Switched Path (LSP) using TLVs carried in LSP Ping [RFC4379]. In particular it specifies the mechanisms necessary to establish MPLS-TP OAM entities at the maintenance points for monitoring and performing measurements on an LSP, as well as defining information elements and procedures to configure proactive MPLS-TP OAM functions running between LERs. Initialization and control of on-demand MPLS-TP OAM functions are expected to be carried out by directly accessing network nodes via a management interface; hence configuration and control of on-demand OAM functions are out-of-scope for this document.

The Transport Profile of MPLS must, by definition [RFC5654], be capable of operating without a control plane. Therefore there are several options for configuring MPLS-TP OAM, without a control plane by either using an NMS or LSP Ping, or with a control plane using signaling protocols RSVP Traffic engineering (RSVP-TE) [RFC3209] and/or Targeted LDP [RFC5036].

Proactive MPLS-TP OAM is performed by set of protocols, Bi-directional Forwarding Detection (BFD) [RFC6428] for Continuity Check/Connectivity Verification, the delay measurement protocol (DM) [RFC6374], [RFC6375] for delay and delay variation (jitter) measurements, and the loss measurement (LM) protocol [RFC6374], [RFC6375] for packet loss and throughput measurements. Additionally, there is a number of Fault Management Signals that can be configured [RFC6427].

BFD is a protocol that provides low-overhead, fast detection of failures in the path between two forwarding engines, including the interfaces, data link(s), and, to the extent possible, the forwarding engines themselves. BFD can be used to detect the continuity and mis-connection defects of MPLS-TP point-to-point and might also be extended to support point-to-multipoint label switched paths (LSPs).

The delay and loss measurements protocols [RFC6374] and [RFC6375] use a simple query/response model for performing both uni- and bi-directional measurements that allow the originating node to measure packet loss and delay in forward or forward and reverse directions.

By timestamping and/or writing current packet counters to the measurement packets (four times, Transmit and Receive in both directions), current delays and packet losses can be calculated. By performing successive delay measurements, the delay and/or inter-packet delay variation (jitter) can be calculated. Current throughput can be calculated from the packet loss measurements by dividing the number of packets sent/received with the time it took to perform the measurement, given by the timestamp in LM header. Combined with a packet generator the throughput measurement can be used to measure the maximum capacity of a particular LSP. It should be noted that this document does not specify how to configure on-demand throughput estimates based on saturating the connection as defined in [RFC6371]. Rather, only how to enable the estimation of the current throughput based on loss measurements.

1.1. Conventions used in this document

1.1.1. Terminology

BFD - Bidirectional Forwarding Detection

DM - Delay Measurement

FMS - Fault Management Signal

G-ACh - Generic Associated Channel

LSP - Label Switched Path

LM - Loss Measurement

MEP - Maintenance Entity Group End Point

MPLS - Multi-Protocol Label Switching

MPLS-TP - MPLS Transport Profile

NMS - Network management System

PM - Performance Measurement

RSVP-TE - RSVP Traffic Engineering

TC - Traffic Class

1.1.2. 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].

2. Theory of Operations

2.1. MPLS OAM Configuration Operation Overview

The MPLS-TP OAM tool set is described in the [RFC6669].

LSP Ping, or alternatively RSVP-TE [RFC7487], can be used to simply enable the different OAM functions, by setting the corresponding flags in the MPLS OAM Functions TLV (refer to Section 2.2). For a more detailed configuration, one may include sub-TLVs for the different OAM functions in order to specify various parameters in detail.

Typically intermediate nodes simply forward OAM configuration TLVs to the end-node without any processing or modification. At least one exception to this is if the FMS sub-TLV (refer to Section 2.2.9) is present. This sub-TLV MUST be examined even by intermediate nodes that support this extension. The sub-TLV MAY be present if a flag is set in the MPLS OAM Functions TLV.

2.1.1. Configuration of BFD Sessions

For this specification, BFD MUST run in either one of the two modes:

- Asynchronous mode, where both sides are in active mode

- Unidirectional mode

In the simplest scenario, LSP Ping [RFC5884], or alternatively RSVP-TE [RFC7487], is used only to bootstrap a BFD session for an LSP, without any timer negotiation.

Timer negotiation can be performed either in subsequent BFD control messages (in this case the operation is similar to LSP Ping based bootstrapping described in [RFC5884]) or directly in the LSP-Ping configuration messages.

When BFD Control packets are transported in the ACH encapsulation, they are not protected by any end-to-end checksum, only lower-layers are providing error detection/correction. A single bit error, e.g. a flipped bit in the BFD State field could cause the receiving end to wrongly conclude that the link is down and in turn trigger protection

switching. To prevent this from happening, the BFD Configuration sub-TLV (refer to Section 2.2.1) has an Integrity flag that when set enables BFD Authentication using Keyed SHA1 with an empty key (all 0s) [RFC5880]. This would make every BFD Control packet carry an SHA1 hash of itself that can be used to detect errors.

If BFD Authentication using a pre-shared key/password is desired (i.e. authentication and not only error detection), the BFD Authentication sub-TLV (refer to Section 2.2.4) MUST be included in the BFD Configuration sub-TLV. The BFD Authentication sub-TLV is used to specify which authentication method that should be used and which pre-shared key/ password that should be used for this particular session. How the key exchange is performed is out of scope of this document.

2.1.2. Configuration of Performance Monitoring

It is possible to configure Performance Monitoring functionalities such as Loss, Delay, Delay/Interpacket Delay variation (jitter), and Throughput as described in [RFC6374].

When configuring Performance Monitoring functionalities, it is possible to choose either the default configuration, by only setting the respective flags in the MPLS OAM functions TLV, or a customized configuration. To customize the configuration, one would set the respective flags in the MPLS OAM functions TLV and include the respective Loss and/or Delay sub-TLVs.

By setting the PM Loss flag in the MPLS OAM Functions TLV and including the PM Loss sub-TLV (refer to Section 2.2.7) one can configure the measurement interval and loss threshold values for triggering protection.

Delay measurements are configured by setting the PM Delay flag in the MPLS OAM Functions TLV and including the PM Delay sub-TLV (refer to Section 2.2.8) one can configure the measurement interval and the delay threshold values for triggering protection.

2.1.3. Configuration of Fault Management Signals

To configure Fault Management Signals (FMS) and their refresh time, the FMS flag in the MPLS OAM Functions TLV MUST be set and the FMS sub-TLV MUST be included. When configuring FMS, an implementation can enable the default configuration by setting the FMS flag in the OAM Function Flags sub-TLV. In order to modify the default configuration, the MPLS OAM FMS sub-TLV MUST be included.

If an intermediate point is meant to originate fault management signal messages, this means that such an intermediate point is associated with a Server MEP through a co-located MPLS-TP client/ server adaptation function, and the Fault Management subscription flag in the MPLS OAM FMS sub-TLV has been set as indication of the request to create the association at each intermediate node of the client LSP. The corresponding Server MEP needs to be configured by its own LSP-ping session or, alternatively, via a Network Management system (NMS) or RSVP-TE.

2.2. MPLS OAM Functions TLV

The MPLS OAM Functions TLV presented in Figure 1 is carried as a TLV of the MPLS Echo Request/Reply messages [RFC4379].

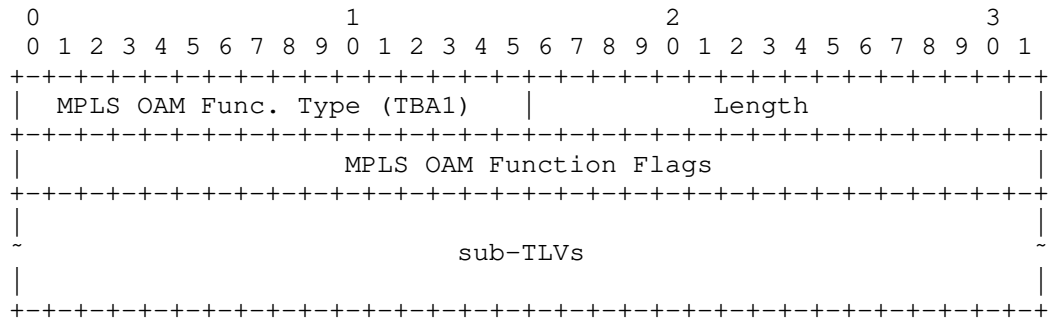


Figure 1: MPLS OAM Functions TLV format

The MPLS OAM Functions TLV contains MPLS OAM Function Flags field. The MPLS OAM Function Flags indicates which OAM functions should be activated as well as OAM function specific sub-TLVs with configuration parameters for the particular function.

Type: indicates the MPLS OAM Functions TLV Section 4.

Length: the length of the MPLS OAM Function Flags field including the total length of the sub-TLVs in octets.

MPLS OAM Function Flags: a bitmap numbered from left to right as shown in the Figure 2. These flags are managed by IANA (refer to Section 4.2). Flags defined in this document are presented in Table 2. Undefined flags MUST be set to zero and unknown flags MUST be ignored. The flags indicate what OAM is being configured and direct the presence of optional sub-TLVs as set out below.

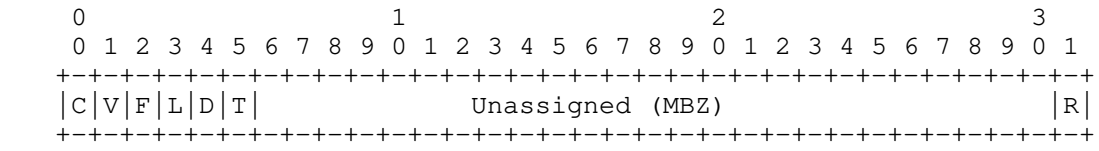


Figure 2: MPLS OAM Function Flags format

Sub-TLVs corresponding to the different flags are as follows. No meaning should be attached to the order of sub-TLVs.

- If a flag in the MPLS OAM Function Flags is set and the corresponding sub-TLVs listed below is absent, then this MPLS OAM function MUST be initialized according to its default settings. Default settings of MPLS OAM functions are outside the scope of this document.

- If any sub-TLV is present without the corresponding flag being set, the sub-TLV SHOULD be ignored.

- BFD Configuration sub-TLV, which MUST be included if either the CC, the CV or both MPLS OAM Function flags being set in the MPLS OAM Functions TLV .

- Performance Monitoring sub-TLV MUST be used to carry PM Loss sub-TLV and/or PM Delay sub-TLV. If neither one of these sub-TLVs is present then Performance Monitoring sub-TLV SHOULD NOT be included. Empty, i.e. no enclosed sub-TLVs, Performance Monitoring sub-TLV SHOULD be ignored.

- PM Loss sub-TLV MAY be included if the PM/Loss OAM Function flag is set. If the "PM Loss sub-TLV" is not included, default configuration values are used. Such sub-TLV MAY also be included in case the Throughput function flag is set and there is the need to specify a measurement interval different from the default ones. In fact, the throughput measurement makes use of the same tool as the loss measurement, hence the same TLV is used.

- PM Delay sub-TLV MAY be included if the PM/Delay OAM Function flag is set. If the "PM Delay sub-TLV" is not included, default configuration values are used.

- FMS sub-TLV, which MAY be included if the FMS OAM Function flag is set. If the "FMS sub-TLV" is not included, default configuration values are used.

If all flags in the MPLS OAM Function Flags field have the same value of zero, that MUST be interpreted as the MPLS OAM Functions TLV not

present in the MPLS Echo Request. If more than one MPLS OAM Functions TLV is present in the MPLS Echo request packet, then the first TLV SHOULD be processed and the rest be ignored. Any parsing error within nested sub-TLVs that is not specified in Section 3 SHOULD be treated as described in [RFC4379].

2.2.1. BFD Configuration Sub-TLV

The BFD Configuration sub-TLV, depicted in Figure 3, is defined for BFD OAM specific configuration parameters. The "BFD Configuration sub-TLV" is carried as a sub-TLV of the "OAM Functions TLV".

This TLV accommodates generic BFD OAM information and carries sub-TLVs.

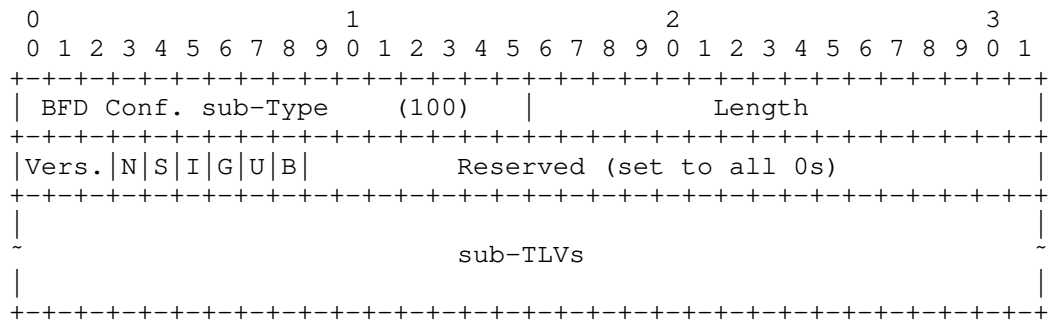


Figure 3: BFD Configuration sub-TLV format

Sub-type: indicates a new sub-type, the BFD Configuration sub-TLV (value 100).

Length: indicates the length of the Value field in octets.

Version: identifies the BFD protocol version. If a node does not support a specific BFD version an error must be generated: "OAM Problem/Unsupported OAM Version".

BFD Negotiation (N): If set timer negotiation/re-negotiation via BFD Control Messages is enabled, when cleared it is disabled and timer configuration is achieved using Negotiation Timer Parameters sub-TLV as described in Section 2.2.3.

Symmetric session (S): If set the BFD session MUST use symmetric timing values. If cleared the BFD session MAY use any timing values either negotiated or explicitly configured.

Integrity (I): If set BFD Authentication MUST be enabled. If the BFD Configuration sub-TLV does not include a BFD Authentication sub-TLV the authentication MUST use Keyed SHA1 with an empty pre-shared key (all 0s). If the egress LSR does not support BFD Authentication an error MUST be generated: "OAM Problem/BFD Authentication unsupported". If the Integrity flag is clear, then Authentication MUST NOT be used.

Encapsulation Capability (G): if set, it shows the capability of encapsulating BFD messages into G-ACh channel. If both the G bit and U bit are set, configuration gives precedence to the G bit.

Encapsulation Capability (U): if set, it shows the capability of encapsulating BFD messages into IP/UDP packets. If both the G bit and U bit are set, configuration gives precedence to the G bit.

If the egress LSR does not support any of the ingress LSR Encapsulation Capabilities an error MUST be generated: "OAM Problem/ Unsupported BFD Encapsulation format".

Bidirectional (B): if set, it configures BFD in the Bidirectional mode. If it is not set it configures BFD in unidirectional mode. In the second case, the source node does not expect any Discriminator values back from the destination node.

Reserved: Reserved for future specification and set to 0 on transmission and ignored when received.

The BFD Configuration sub-TLV MUST include the following sub-TLVs in the MPLS Echo Request message:

- Local Discriminator sub-TLV, if B flag is set in the MPLS Echo Request;

- Negotiation Timer Parameters sub-TLV if the N flag is cleared.

The BFD Configuration sub-TLV MUST include the following sub-TLVs in the MPLS Echo Reply message:

- Local Discriminator sub-TLV;

- Negotiation Timer Parameters sub-TLV if:

  - the N and S flags are cleared, or if:

  - the N flag is cleared and the S flag is set, and the Negotiation Timer Parameters sub-TLV received by the egress contains unsupported values. In this case an updated

Negotiation Timer Parameters sub-TLV, containing values supported by the egress node [RFC7419], is returned to the ingress.

2.2.2. Local Discriminator Sub-TLV

The Local Discriminator sub-TLV is carried as a sub-TLV of the "BFD Configuration sub-TLV" and is depicted in Figure 4.

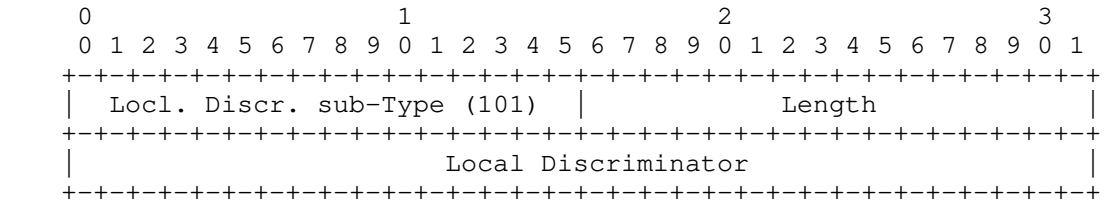


Figure 4: Local Discriminator sub-TLV format

Type: indicates a new type, the "Local Discriminator sub-TLV" (value 101).

Length: indicates the length of the Value field in octets . (4)

Local Discriminator: A nonzero discriminator value that is unique in the context of the transmitting system that generates it. It is used to demultiplex multiple BFD sessions between the same pair of systems.

2.2.3. Negotiation Timer Parameters Sub-TLV

The Negotiation Timer Parameters sub-TLV is carried as a sub-TLV of the BFD Configuration sub-TLV and is depicted in Figure 5.

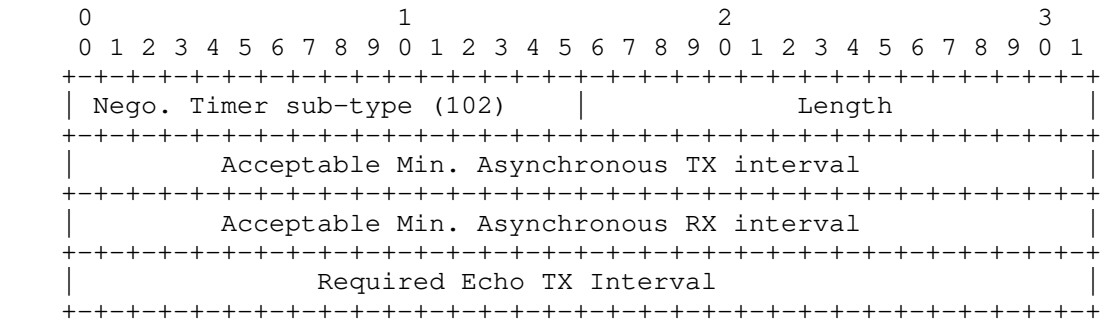


Figure 5: Negotiation Timer Parameters sub-TLV format

Sub-type: indicates a new sub-type, the Negotiation Timer Parameters sub-TLV (value 102).

Length: indicates the length of the Value field in octets (12).

Acceptable Min. Asynchronous TX interval: in case of S (symmetric) flag set in the BFD Configuration sub-TLV, defined in Section 2.2.1, it expresses the desired time interval (in microseconds) at which the ingress LER intends to both transmit and receive BFD periodic control packets. If the receiving edge LSR cannot support such value, it SHOULD reply with an interval greater than the one proposed.

In case of S (symmetric) flag cleared in the BFD Configuration sub-TLV, this field expresses the desired time interval (in microseconds) at which a edge LSR intends to transmit BFD periodic control packets in its transmitting direction.

Acceptable Min. Asynchronous RX interval: in case of S (symmetric) flag set in the BFD Configuration sub-TLV, Figure 3, this field MUST be equal to Acceptable Min. Asynchronous TX interval and has no additional meaning respect to the one described for "Acceptable Min. Asynchronous TX interval".

In case of S (symmetric) flag cleared in the BFD Configuration sub-TLV, it expresses the minimum time interval (in microseconds) at which edge LSRs can receive BFD periodic control packets. In case this value is greater than the value of Acceptable Min. Asynchronous TX interval received from the other edge LSR, such edge LSR MUST adopt the interval expressed in this Acceptable Min. Asynchronous RX interval.

Required Echo TX Interval: the minimum interval (in microseconds) between received BFD Echo packets that this system is capable of supporting, less any jitter applied by the sender as described in [RFC5880] sect. 6.8.9. This value is also an indication for the receiving system of the minimum interval between transmitted BFD Echo packets. If this value is zero, the transmitting system does not support the receipt of BFD Echo packets. If the receiving system cannot support this value the "Unsupported BFD TX Echo rate interval" error MUST be generated. By default the value is set to 0.

2.2.4. BFD Authentication Sub-TLV

The "BFD Authentication sub-TLV" is carried as a sub-TLV of the "BFD Configuration sub-TLV" and is depicted in Figure 6.

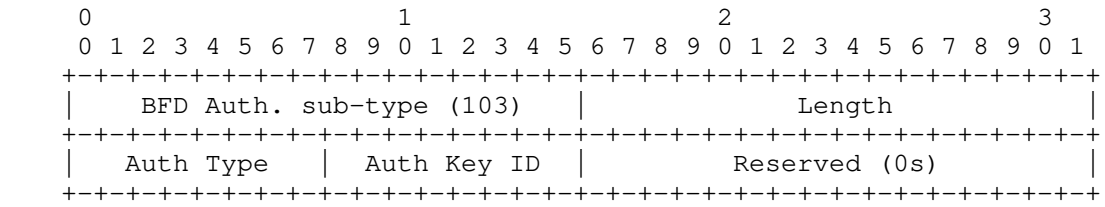


Figure 6: BFD Authentication sub-TLV format

Sub-type: indicates a new type, the BFD Authentication sub-TLV (value 103).

Length: indicates the length of the Value field in octets (4).

Auth Type: indicates which type of authentication to use. The same values as are defined in section 4.1 of [RFC5880] are used. Simple Password SHOULD NOT be used if other authentication types are available.

Auth Key ID: indicates which authentication key or password (depending on Auth Type) should be used. How the key exchange is performed is out of scope of this document. If the egress LSR does not support this Auth Key ID an "OAM Problem/Mismatch of BFD Authentication Key ID" error MUST be generated.

Reserved: Reserved for future specification and set to 0 on transmission and ignored when received.

An implementation MAY change mode of authentication if an operator re-evaluates the security situation in and around the administrative domain. If BFD Authentication sub-TLV used for a BFD session in Up state, then the Sender of the MPLS LSP Echo Request SHOULD ensure that old and new modes of authentication, i.e. combination of Auth.Type and Auth. Key ID, are used to send and receive BFD control packets, until the Sender can confirm that its peer has switched to the new authentication.

2.2.5. Traffic Class Sub-TLV

The Traffic Class sub-TLV is carried as a sub-TLV of the "BFD Configuration sub-TLV" and "Fault Management Signal sub-TLV" Section 2.2.9 and is depicted in Figure 7.

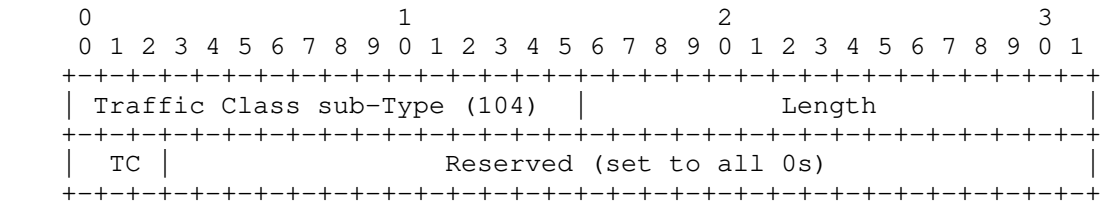


Figure 7: Traffic Class sub-TLV format

Type: indicates a new type, the "Traffic Class sub-TLV" (value 104).

Length: indicates the length of the Value field in octets . (4)

TC: Identifies the Traffic Class (TC) [RFC5462] for periodic continuity monitoring messages or packets with fault management information.

If the TC sub-TLV is present, then the sender of any periodic continuity monitoring messages or packets with fault management information on the LSP, with a FEC that corresponds to the FEC for which fault detection is being performed, MUST use the value contained in the TC field of the sub-TLV as the value of the TC field in the top label stack entry of the MPLS label stack. If the TC sub-TLV is absent from either "BFD Configuration sub-TLV" or "Fault Management Signal sub-TLV", then selection of the TC value is local decision.

2.2.6. Performance Measurement Sub-TLV

If the MPLS OAM Functions TLV has any of the L (Loss), D (Delay) and T (Throughput) flag set, the Performance Measurement sub-TLV MUST be present. Failure to include the correct sub-TLVs MUST result in an "OAM Problem/ Configuration Error" error being generated.

The Performance Measurement sub-TLV provides the configuration information mentioned in Section 7 of [RFC6374]. It includes support for the configuration of quality thresholds and, as described in [RFC6374], "the crossing of which will trigger warnings or alarms, and result in reporting and exception notification will be integrated into the system-wide network management and reporting framework."

In case the values need to be different than the default ones, the Performance Measurement sub-TLV MAY include the following sub-TLVs:

- PM Loss sub-TLV if the L flag is set in the MPLS OAM Functions TLV;

- PM Delay sub-TLV if the D flag is set in the MPLS OAM Functions TLV.

The Performance Measurement sub-TLV depicted in Figure 8 is carried as a sub-TLV of the MPLS OAM Functions TLV.

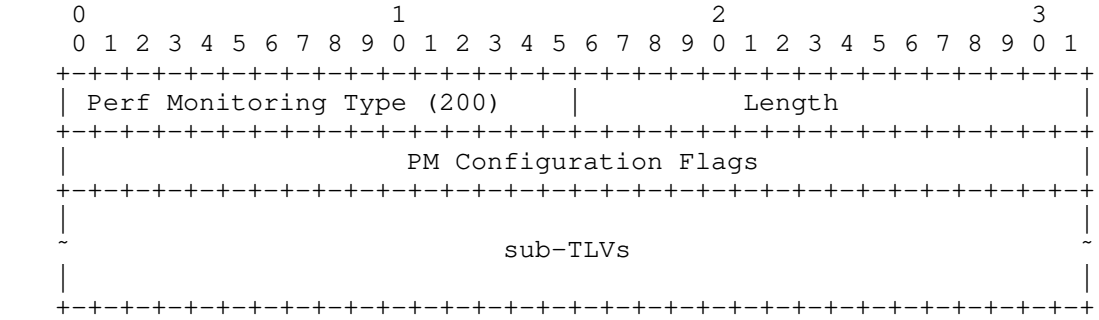


Figure 8: Performance Measurement sub-TLV format

Sub-type: indicates a new sub-type, the Performance Management sub-TLV" (value 200).

Length: indicates the length of the Value field in octets, including PM Configuration Flags and optional sub-TLVs.

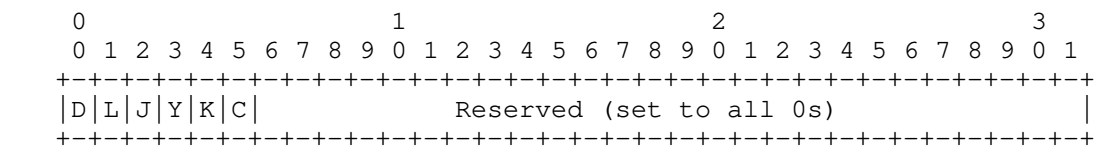


Figure 9: Performance Measurement sub-TLV format

PM Configuration Flags, format is presented in Figure 9, for the specific function description please refer to [RFC6374]:

- D: Delay inferred/direct (0=INFERRED, 1=DIRECT). If the egress LSR does not support specified mode an "OAM Problem/Unsupported Delay Mode" error MUST be generated.

- L: Loss inferred/direct (0=INFERRED, 1=DIRECT). If the egress LSR does not support specified mode an "OAM Problem/Unsupported Loss Mode" error MUST be generated.

- J: Delay variation/jitter (1=ACTIVE, 0=NOT ACTIVE). If the egress LSR does not support Delay variation measurements and the J

flag is set, an "OAM Problem/Delay variation unsupported" error MUST be generated.

- Y: Dyadic (1=ACTIVE, 0=NOT ACTIVE). If the egress LSR does not support Dyadic mode and the Y flag is set, an "OAM Problem/Dyadic mode unsupported" error MUST be generated.

- K: Loopback (1=ACTIVE, 0=NOT ACTIVE). If the egress LSR does not support Loopback mode and the K flag is set, an "OAM Problem/ Loopback mode unsupported" error MUST be generated.

- C: Combined (1=ACTIVE, 0=NOT ACTIVE). If the egress LSR does not support Combined mode and the C flag is set, an "OAM Problem/ Combined mode unsupported" error MUST be generated.

Reserved: Reserved for future specification and set to 0 on transmission and ignored when received.

2.2.7. PM Loss Measurement Sub-TLV

The PM Loss Measurement sub-TLV depicted in Figure 10 is carried as a sub-TLV of the Performance Measurement sub-TLV.

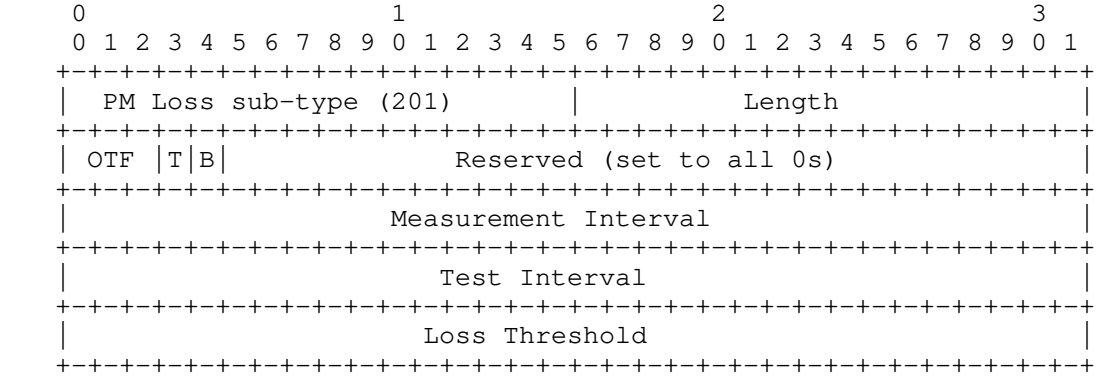


Figure 10: PM Loss Measurement sub-TLV format

Sub-type: indicates a new sub-type, the PM Loss Measurement sub-TLV (value 201).

Length: indicates the length of the Value field in octets (16).

OTF: Origin Timestamp Format of the Origin Timestamp field described in [RFC6374]. By default it is set to IEEE 1588 version 1. If the egress LSR cannot support this value an "OAM Problem/Unsupported Timestamp Format" error MUST be generated.

Configuration Flags, please refer to [RFC6374] for further details:

- T: Traffic-class-specific measurement indicator. Set to 1 when the measurement operation is scoped to packets of a particular traffic class (DSCP value), and 0 otherwise. When set to 1, the DS field of the message indicates the measured traffic class. By default it is set to 1.

- B: Octet (byte) count. When set to 1, indicates that the Counter 1-4 fields represent octet counts. When set to 0, indicates that the Counter 1-4 fields represent packet counts. By default it is set to 0.

Reserved: Reserved for future specification and set to 0 on transmission and ignored when received.

Measurement Interval: the time interval (in milliseconds) at which Loss Measurement query messages MUST be sent on both directions. If the edge LSR receiving the Path message cannot support such value, it SHOULD reply with a higher interval. By default it is set to (100) as per [RFC6375].

Test Interval: test messages interval in milliseconds as described in [RFC6374]. By default it is set to (10) as per [RFC6375].

Loss Threshold: the threshold value of measured lost packets per measurement over which action(s) SHOULD be triggered.

2.2.8. PM Delay Measurement Sub-TLV

The "PM Delay Measurement sub-TLV" depicted in Figure 11 is carried as a sub-TLV of the Performance Monitoring sub-TLV.

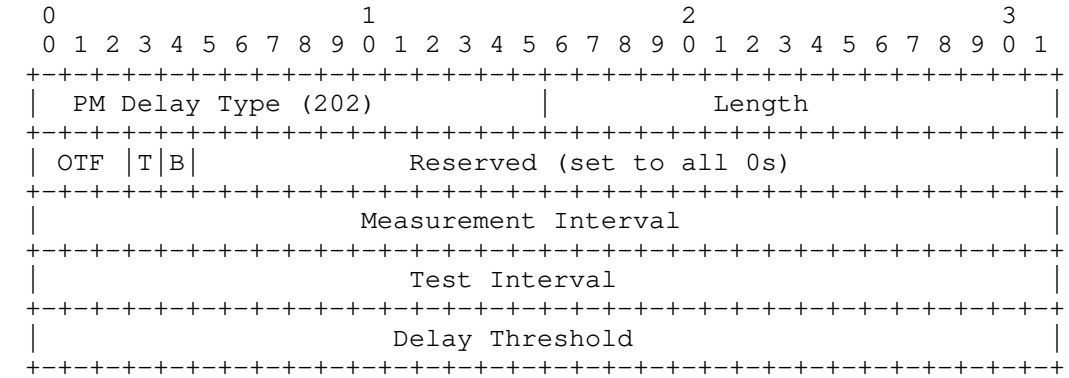


Figure 11: PM Delay Measurement sub-TLV format

Sub-type: indicates a new sub-type, the "PM Delay Measurement sub-TLV" (value 202).

Length: indicates the length of the Value field in octets (16).

OTF: Origin Timestamp Format of the Origin Timestamp field described in [RFC6374]. By default it is set to IEEE 1588 version 1. If the egress LSR cannot support this value, an "OAM Problem/Unsupported Timestamp Format" error MUST be generated.

Configuration Flags, please refer to [RFC6374] for further details:

- T: Traffic-class-specific measurement indicator. Set to 1 when the measurement operation is scoped to packets of a particular traffic class (DSCP value), and 0 otherwise. When set to 1, the DS field of the message indicates the measured traffic class. By default it is set to 1.

- B: Octet (byte) count. When set to 1, indicates that the Counter 1-4 fields represent octet counts. When set to 0, indicates that the Counter 1-4 fields represent packet counts. By default it is set to 0.

Reserved: Reserved for future specification and set to 0 on transmission and ignored when received.

Measurement Interval: the time interval (in milliseconds) at which Delay Measurement query messages MUST be sent on both directions. If the edge LSR receiving the Path message cannot support such value, it can reply with a higher interval. By default it is set to (1000) as per [RFC6375].

Test Interval: test messages interval (in milliseconds) as described in [RFC6374]. By default it is set to (10) as per [RFC6375].

Delay Threshold: the threshold value of measured two-way delay (in milliseconds) over which action(s) SHOULD be triggered.

2.2.9. Fault Management Signal Sub-TLV

The FMS sub-TLV depicted in Figure 12 is carried as a sub-TLV of the MPLS OAM Configuration sub-TLV. When both working and protection paths are configured, both LSPs SHOULD be configured with identical settings of the E flag, T flag, and the refresh timer. An implementation MAY configure the working and protection LSPs with different settings of these fields in case of 1:N protection.

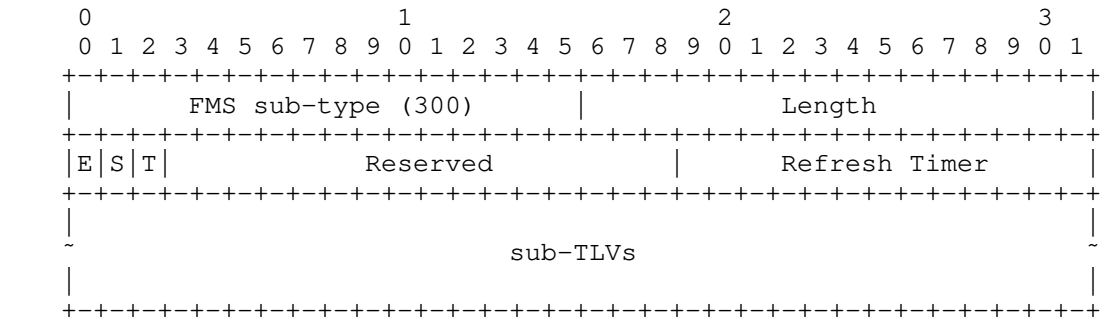


Figure 12: Fault Management Signal sub-TLV format

Sub-type: indicates a new sub-type, the FMS sub-TLV (value 300).

Length: indicates the length of the Value field in octets.

FMS Signal Flags are used to enable the FMS signals at end point MEPs and the Server MEPs of the links over which the LSP is forwarded. In this document only the S flag pertains to Server MEPs.

The following flags are defined:

- E: Enable Alarm Indication Signal (AIS) and Lock Report (LKR) signaling as described in [RFC6427]. Default value is 1 (enabled). If the egress MEP does not support FMS signal generation, an "OAM Problem/Fault management signaling unsupported" error MUST be generated.

- S: Indicate to a server MEP that it should transmit AIS and LKR signals on the client LSP. Default value is 0 (disabled). If a Server MEP which is capable of generating FMS messages is for some reason unable to do so for the LSP being signaled, an "OAM Problem/Unable to create fault management association" error MUST be generated.

- T: Set timer value, enabled the configuration of a specific timer value. Default value is 0 (disabled).

- Remaining bits: Reserved for future specification and set to 0.

Refresh Timer: indicates the refresh timer of fault indication messages, in seconds. The value MUST be between 1 to 20 seconds as specified for the Refresh Timer field in [RFC6427]. If the edge LSR receiving the Path message cannot support the value it SHOULD reply with a higher timer value.

FMS sub-TLV MAY include Traffic Class sub-TLV Section 2.2.5. If TC sub-TLV is present, the value of the TC field MUST be used as the value of the TC field of an MPLS label stack entry for FMS messages. If the TC sub-TLV is absent, then selection of the TC value is local decision.

2.2.10. Source MEP-ID Sub-TLV

The Source MEP-ID sub-TLV depicted in Figure 13 is carried as a sub-TLV of the MPLS OAM Functions TLV.

Note that support of ITU IDs is out-of-scope.

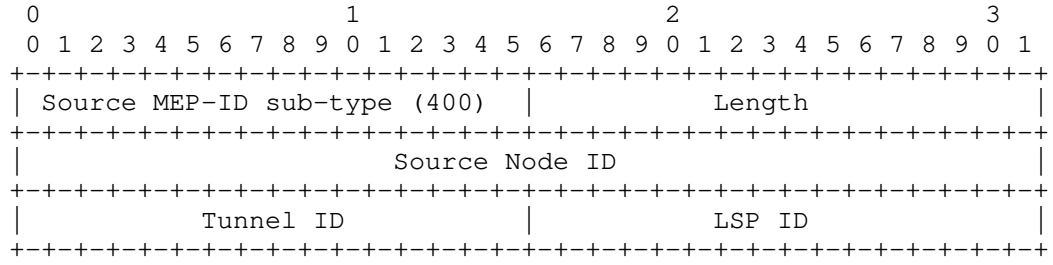


Figure 13: Source MEP-ID sub-TLV format

Sub-type: indicates a new sub-type, the Source MEP-ID sub-TLV (value 400).

Length: indicates the length of the Value field in octets (8).

Source Node ID: 32-bit node identifier as defined in [RFC6370].

Tunnel ID: a 16-bit unsigned integer unique to the node as defined in [RFC6370].

LSP ID: a 16-bit unsigned integer unique within the Tunnel_ID as defined in [RFC6370].

3. Summary of MPLS OAM Configuration Errors

This is the summary of Return Codes [RFC4379] defined in this document:

- If an egress LSR does not support the specified BFD version, an error MUST be generated: "OAM Problem/Unsupported BFD Version".

- If an egress LSR does not support the specified BFD Encapsulation format, an error MUST be generated: "OAM Problem/ Unsupported BFD Encapsulation format".

- If an egress LSR does not support BFD Authentication, and it is requested, an error MUST be generated: "OAM Problem/BFD Authentication unsupported".

- If an egress LSR does not support the specified BFD Authentication Type, an error MUST be generated: "OAM Problem/ Unsupported BFD Authentication Type".

- If an egress LSR is not able to use the specified Authentication Key ID, an error MUST be generated: "OAM Problem/Mismatch of BFD Authentication Key ID".

- If an egress LSR does not support the specified Timestamp Format, an error MUST be generated: "OAM Problem/Unsupported Timestamp Format".

- If an egress LSR does not support specified Delay mode, an "OAM Problem/Unsupported Delay Mode" error MUST be generated.

- If an egress LSR does not support specified Loss mode, an "OAM Problem/Unsupported Loss Mode" error MUST be generated.

- If an egress LSR does not support Delay variation measurements, and it is requested, an "OAM Problem/Delay variation unsupported" error MUST be generated.

- If an egress LSR does not support Dyadic mode, and it is requested, an "OAM Problem/Dyadic mode unsupported" error MUST be generated.

- If an egress LSR does not support Loopback mode, and it is requested, an "OAM Problem/Loopback mode unsupported" error MUST be generated.

- If an egress LSR does not support Combined mode, and it is requested, an "OAM Problem/Combined mode unsupported" error MUST be generated.

- If an egress LSR does not support Fault Monitoring Signals, and it is requested, an "OAM Problem/Fault management signaling unsupported" error MUST be generated.

- If an intermediate server MEP supports Fault Monitoring Signals but is unable to create an association, when requested to do so,

an "OAM Problem/Unable to create fault management association" error MUST be generated.

Ingress LSR MAY combine multiple MPLS OAM configuration TLVs and sub-TLVs into single MPLS echo request. In case an egress LSR doesn't support any of the requested modes it MUST set the return code to report the first unsupported mode in the list of TLVs and sub-TLVs. And if any of the requested OAM configuration is not supported the egress LSR SHOULD NOT process OAM Configuration TLVs and sub-TLVs listed in the MPLS echo request.

4. IANA Considerations

4.1. TLV and Sub-TLV Allocation

IANA maintains the Multi-Protocol Label Switching (MPLS) Label Switched Paths (LSPs) Ping Parameters registry, and within that registry a sub-registry for TLVs and sub-TLVs.

IANA is requested to allocate a new MPLS OAM Functions TLV from the standards action range (0-16383) and sub-TLVs as follows from sub-registry presented in Table 1, called "Sub-TLVs for TLV [TBA1]".

Registration procedures for Sub-TLVs from ranges 0-16383 and 32768-49161 are by Standards Action, and from ranges 16384-31743 and 49162-64511 are through Specification Required (Experimental RFC Needed).

Type	Sub-type	Value Field	Reference
TBA1		MPLS OAM Functions	This document
	100	BFD Configuration	This document
	101	BFD Local Discriminator	This document
	102	BFD Negotiation Timer Parameters	This document
	103	BFD Authentication	This document
	104	Traffic Class	This document
	200	Performance Measurement	This document
	201	PM Loss Measurement	This document
	202	PM Delay Measurement	This document
	300	Fault Management Signal	This document
	400	Source MEP-ID	This document

Table 1: IANA TLV Type Allocation

4.2. MPLS OAM Function Flags Allocation

IANA is requested to create a new registry called the "MPLS OAM Function Flags" registry . Assignments of bit positions 0 through 31 are via Standards Action. The new registry to be populated as follows.

Bit Position	MPLS OAM Function Flag	Description
0	C	Continuity Check (CC)
1	V	Connectivity Verification (CV)
2	F	Fault Management Signal (FMS)
3	L	Performance Measurement/Loss (PM/Loss)
4	D	Performance Measurement/Delay (PM/Delay)
5	T	Throughput Measurement
6-30		Unassigned (Must be zero)
31		Reserved

Table 2: MPLS OAM Function Flags

4.3. OAM Configuration Errors

IANA maintains a registry "Multi-Protocol Label Switching (MPLS) Label Switched Paths (LSPs) Ping Parameters" registry, and within that registry a sub-registry "Return Codes".

IANA is requested to assign new Return Codes from the Standards Action range (0-191) as follows:

Error Value Sub-codes	Description	Reference
TBA3	OAM Problem/Unsupported BFD Version	This document
TBA4	OAM Problem/Unsupported BFD Encapsulation format	This document
TBA5	OAM Problem/Unsupported BFD Authentication Type	This document
TBA6	OAM Problem/Mismatch of BFD Authentication Key ID	This document
TBA7	OAM Problem/Unsupported Timestamp Format	This document
TBA8	OAM Problem/Unsupported Delay Mode	This document
TBA9	OAM Problem/Unsupported Loss Mode	This document
TBA10	OAM Problem/Delay variation unsupported	This document
TBA11	OAM Problem/Dyadic mode unsupported	This document
TBA12	OAM Problem/Loopback mode unsupported	This document
TBA13	OAM Problem/Combined mode unsupported	This document
TBA14	OAM Problem/Fault management signaling unsupported	This document
TBA15	OAM Problem/Unable to create fault management association	This document

Table 3: IANA Return Codes Allocation

5. Security Considerations

The signaling of OAM related parameters and the automatic establishment of OAM entities introduces additional security considerations to those discussed in [RFC4379]. In particular, a network element could be overloaded if an attacker were to request high frequency liveliness monitoring of a large number of LSPs, targeting a single network element. Implementations must be made cognizant of available OAM resources and MAY refuse new OAM configurations that would overload a node. Additionally, policies to manage OAM resources may be used to provide some fairness in OAM resource distribution among monitored LSPs.

Security of OAM protocols configured with extensions to LSP Ping described in this document are discussed in [RFC5880], [RFC5884], [RFC6374], [RFC6427], and [RFC6428].

In order that the configuration of OAM functionality can be achieved securely through the techniques described in this document, security mechanisms must already be in place and operational for LSP Ping. Thus the exchange of security parameters (such as keys) for use in securing OAM is outside the scope of this document and is assumed to use an off-line mechanism or an established secure key-exchange protocol.

Additional discussion of security for MPLS protocols can be found in [RFC5920].

6. Acknowledgements

The authors would like to thank Nobo Akiya, David Allan and Adrian Farrel for their thorough reviews and insightful comments.

7. References

7.1. Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <http://www.rfc-editor.org/info/rfc2119>.

[RFC4379] Kompella, K. and G. Swallow, "Detecting Multi-Protocol Label Switched (MPLS) Data Plane Failures", RFC 4379, DOI 10.17487/RFC4379, February 2006, <http://www.rfc-editor.org/info/rfc4379>.

[RFC5654] Niven-Jenkins, B., Ed., Brungard, D., Ed., Betts, M., Ed., Sprecher, N., and S. Ueno, "Requirements of an MPLS Transport Profile", RFC 5654, DOI 10.17487/RFC5654, September 2009, <http://www.rfc-editor.org/info/rfc5654>.

[RFC5880] Katz, D. and D. Ward, "Bidirectional Forwarding Detection (BFD)", RFC 5880, DOI 10.17487/RFC5880, June 2010, <http://www.rfc-editor.org/info/rfc5880>.

[RFC5884] Aggarwal, R., Kompella, K., Nadeau, T., and G. Swallow, "Bidirectional Forwarding Detection (BFD) for MPLS Label Switched Paths (LSPs)", RFC 5884, DOI 10.17487/RFC5884, June 2010, <http://www.rfc-editor.org/info/rfc5884>.

[RFC6370] Bocci, M., Swallow, G., and E. Gray, "MPLS Transport Profile (MPLS-TP) Identifiers", RFC 6370, DOI 10.17487/RFC6370, September 2011, <http://www.rfc-editor.org/info/rfc6370>.

[RFC6374] Frost, D. and S. Bryant, "Packet Loss and Delay Measurement for MPLS Networks", RFC 6374, DOI 10.17487/RFC6374, September 2011, <http://www.rfc-editor.org/info/rfc6374>.

[RFC6427] Swallow, G., Ed., Fulignoli, A., Ed., Vigoureux, M., Ed., Boutros, S., and D. Ward, "MPLS Fault Management Operations, Administration, and Maintenance (OAM)", RFC 6427, DOI 10.17487/RFC6427, November 2011, <http://www.rfc-editor.org/info/rfc6427>.

[RFC6428] Allan, D., Ed., Swallow Ed., G., and J. Drake Ed., "Proactive Connectivity Verification, Continuity Check, and Remote Defect Indication for the MPLS Transport Profile", RFC 6428, DOI 10.17487/RFC6428, November 2011, <http://www.rfc-editor.org/info/rfc6428>.

7.2. Informative References

[RFC3209] Awduche, D., Berger, L., Gan, D., Li, T., Srinivasan, V., and G. Swallow, "RSVP-TE: Extensions to RSVP for LSP Tunnels", RFC 3209, DOI 10.17487/RFC3209, December 2001, <http://www.rfc-editor.org/info/rfc3209>.

[RFC5036] Andersson, L., Ed., Minei, I., Ed., and B. Thomas, Ed., "LDP Specification", RFC 5036, DOI 10.17487/RFC5036, October 2007, <http://www.rfc-editor.org/info/rfc5036>.

[RFC5462] Andersson, L. and R. Asati, "Multiprotocol Label Switching (MPLS) Label Stack Entry: "EXP" Field Renamed to "Traffic Class" Field", RFC 5462, DOI 10.17487/RFC5462, February 2009, <http://www.rfc-editor.org/info/rfc5462>.

[RFC5860] Vigoureux, M., Ed., Ward, D., Ed., and M. Betts, Ed., "Requirements for Operations, Administration, and Maintenance (OAM) in MPLS Transport Networks", RFC 5860, DOI 10.17487/RFC5860, May 2010, <http://www.rfc-editor.org/info/rfc5860>.

[RFC5920] Fang, L., Ed., "Security Framework for MPLS and GMPLS Networks", RFC 5920, DOI 10.17487/RFC5920, July 2010, <http://www.rfc-editor.org/info/rfc5920>.

[RFC6371] Busi, I., Ed. and D. Allan, Ed., "Operations, Administration, and Maintenance Framework for MPLS-Based Transport Networks", RFC 6371, DOI 10.17487/RFC6371, September 2011, <http://www.rfc-editor.org/info/rfc6371>.

[RFC6375] Frost, D., Ed. and S. Bryant, Ed., "A Packet Loss and Delay Measurement Profile for MPLS-Based Transport Networks", RFC 6375, DOI 10.17487/RFC6375, September 2011, <http://www.rfc-editor.org/info/rfc6375>.

[RFC6669] Sprecher, N. and L. Fang, "An Overview of the Operations, Administration, and Maintenance (OAM) Toolset for MPLS-Based Transport Networks", RFC 6669, DOI 10.17487/RFC6669, July 2012, <http://www.rfc-editor.org/info/rfc6669>.

[RFC7419] Akiya, N., Binderberger, M., and G. Mirsky, "Common Interval Support in Bidirectional Forwarding Detection", RFC 7419, DOI 10.17487/RFC7419, December 2014, <http://www.rfc-editor.org/info/rfc7419>.

[RFC7487] Bellagamba, E., Takacs, A., Mirsky, G., Andersson, L., Skoldstrom, P., and D. Ward, "Configuration of Proactive Operations, Administration, and Maintenance (OAM) Functions for MPLS-Based Transport Networks Using RSVP-TE", RFC 7487, DOI 10.17487/RFC7487, March 2015, <http://www.rfc-editor.org/info/rfc7487>.

Authors' Addresses

Elisa Bellagamba

Email: elisa.bellagamba@gmail.com

Gregory Mirsky
Ericsson

Email: Gregory.Mirsky@ericsson.com

Loa Andersson
Huawei Technologies

Email: loa@mail01.huawei.com

Pontus Skoldstrom
Acreo AB
Electrum 236
Kista 164 40
Sweden

Phone: +46 8 6327731
Email: pontus.skoldstrom@acreo.se

Dave Ward
Cisco

Email: dward@cisco.com

John Drake
Juniper

Email: jdrake@juniper.net

Network Working Group
INTERNET-DRAFT
Intended Status: Standards Track

Sami Boutros
Siva Sivabalan
George Swallow
Shaleen Saxena
Cisco Systems

Vishwas Manral
Hewlett Packard Co.

Sam Aldrin
Huawei Technologies, Inc.

Expires: February 20, 2015 August 19, 2014

Definition of Time-to-Live TLV for LSP-Ping Mechanisms
draft-ietf-mpls-lsp-ping-ttl-tlv-10.txt

Abstract

LSP-Ping is a widely deployed Operation, Administration, and Maintenance (OAM) mechanism in MPLS networks. However, in the present form, this mechanism is inadequate to verify connectivity of a segment of a Multi-Segment PseudoWire (MS-PW) and/or bidirectional co-routed LSP from any node on the path of the MS-PW and/or bidirectional co-routed LSP. This document defines a TLV to address this shortcoming.

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/1id-abstracts.html

The list of Internet-Draft Shadow Directories can be accessed at

http://www.ietf.org/shadow.html

Copyright and License Notice

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

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

Table of Contents

1. Introduction

A MS-PW may span across multiple service provider networks. In order to allow Service Providers (SP) to verify segments of such MS-PW from any node on the path of the MS-PW, any node along the path of the MS-PW, should be able to originate an MPLS Echo Request packet to any other node along the path of the MS-PW and receive the corresponding MPLS Echo Reply. If the originator of the MPLS Echo Request is at the end of a MS-PW, the receiver of the request can send the reply back to the sender without knowing the hop-count distance of the originator. The reply will be intercepted by the originator regardless of the TTL value on the reply packet. But, if the originator is not at the end of the MS-PW, the receiver of the MPLS Echo Request may need to know how many hops away the originator of the MPLS Echo Request is so that it can set the TTL value on the MPLS header for the MPLS Echo Reply to be intercepted at the originator node.

In MPLS networks, for bidirectional co-routed LSPs, if it is desired to verify connectivity from any intermediate node (LSR) on the LSP to the any other LSR on the LSP the receiver may need to know the TTL to send the MPLS Echo Reply with, so as the packet is intercepted by the originator node.

A new optional TTL TLV is defined in this document. This TLV will be added by the originator of the MPLS Echo Request to inform the receiver how many hops away the originator is on the path of the MS-PW or Bidirectional LSP.

This mechanism only works if the MPLS Echo Reply is sent down the co-routed LSP, hence the scope of this TTL TLV is currently limited to MS-PW or Bidirectional co-routed MPLS LSPs. The presence of the TLV implies the use of the return path of the co-routed LSP, if the return path is any other mechanism then the TLV in the MPLS Echo Request MUST be ignored.

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].

LSR: Label Switching Router

MPLS-TP: MPLS Transport Profile

MS-PW: Multi-Segment Pseudowire

PW: Pseudowire

TLV: Type Length Value

TTL: Time To Live

3. Time To Live TLV

3.1. TTL TLV Format

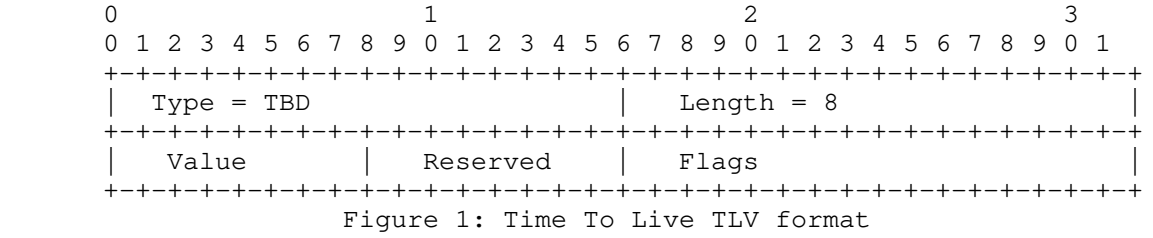


Figure 1: Time To Live TLV format

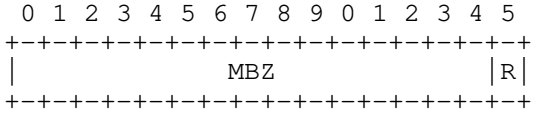
The TTL TLV has the format shown in Figure 1.

Value

The value of the TTL as specified by this TLV

Flags

The Flags field is a bit vector with the following format:



One flag is defined for now, the R flag. The rest of the flags are Reserved - MUST be zero (MBZ) when sending and ignored on receipt.

The R flag (Reply TTL) is set signify that the value is meant to be used as the TTL for the reply packet. Other bits may be defined later to enhance the scope of this TLV.

3.2. Usage

The TTL TLV MAY be included in the MPLS Echo Request by the

originator of the request.

If the TTL TLV is present and the receiver does not understand TTL TLVs, it will simply ignore the TLV, as is the case for all optional TLVs. If the TTL TLV is not present or is not processed by the receiver, any determination of the TTL value used in the MPLS label on the LSP-Ping echo reply is beyond the scope of this document.

If the TTL TLV is present and the receiver understands TTL TLVs, one of the following two conditions apply:

o If the TTL TLV value field is zero, the LSP-Ping echo request packet SHOULD be dropped.

o Otherwise, the receiver MUST use the TTL value specified in the TTL TLV when it creates the MPLS header of the MPLS Echo Reply. The TTL value in the TTL TLV takes precedence over any TTL value determined by other means, such as from the Switching Point TLV in the MS-PW. This precedence will aid the originator of the LSP-Ping echo request in analyzing the return path.

4. Operation

In this section, we explain a use case for the TTL TLV with an MPLS MS-PW.

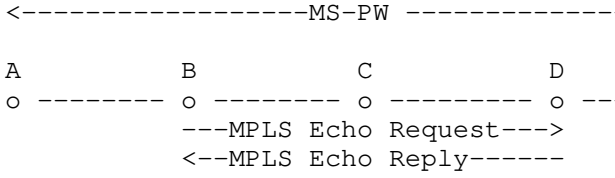


Figure 2: Use-case with MS-PWs

Let us assume a MS-PW going through LSRs A, B, C, D, and E. Furthermore, assume that an operator wants to perform a connectivity check between B and D from B. Thus, an MPLS Echo Request with the TTL TLV is originated from B and sent towards D. The MPLS Echo Request packet contains the FEC of the PW Segment between C and D. The value field of the TTL TLV and the TTL field of the MPLS label are set to 2, the choice of the value 2 will be based on the operator input requesting the MPLS Echo Request or from the optional LDP switching point TLV. The MPLS Echo Request is intercepted at D because of TTL expiry. D detects the TTL TLV in the request, and use the TTL value (i.e., 2) specified in the TLV on the MPLS label of the MPLS Echo Reply. The MPLS Echo Reply will be intercepted by B because of TTL

expiry.

The same operation will apply when we have a co-routed bidirectional LSP, and we want to check connectivity from an intermediate LSR "B" to another LSR "D".

4.1. Traceroute mode

In traceroute mode, the TTL value in the TLV is set to 1 for the first Echo Request, then to 2 for the next, and so on. This is similar to the TTL values used for the label set on the packet.

4.2. Error scenario

It is possible that the MPLS Echo Request packet was intercepted before the intended destination for reason other than label TTL expiry. This could be due network faults, misconfiguration or other reasons. In such cases, if the return TTL is set to the value specified in the TTL TLV then the echo response packet will continue beyond the originating node. This becomes a security issue.

To prevent this, the label TTL value used in the MPLS Echo Reply packet MUST be modified by deducting the incoming label TTL on the received packet from TTL TLV value. If the MPLS Echo Request packet is punted to the CPU before the incoming label TTL is deducted, then another 1 MUST be added. In other words:

Return TTL Value on the MPLS Echo Reply packet = (TTL TLV Value)- (Incoming Label TTL) + 1

5. Security Considerations

This draft allows the setting of the TTL value in the MPLS Label of an MPLS Echo Reply, so that it can be intercepted by an intermediate device. This can cause a device to get a lot of LSP Ping packets which get redirected to the CPU.

However the same is possible even without the changes mentioned in this document. A device should rate limit the LSP ping packets redirected to the CPU so that the CPU is not overwhelmed.

The recommendation in [RFC4379] security section applies, to check the source address of the MPLS Echo Request, however the source address can now be any node along the LSP path.

A faulty transit node changing the TTL TLV value could make the wrong node reply to the MPLS Echo Request, and/or the wrong node to receive

the MPLS Echo Reply. An LSP trace may help identify the faulty transit node.

6. IANA Considerations

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

Time To Live TLV (See Section 3). The value MUST be assigned from the range (32768-49161) of optional TLVs.

IANA is requested to allocate the value 32769.

7. Acknowledgements

The authors would like to thank Greg Mirsky for his comments.

8. References

8.1 Normative References

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

[2] T. Nadeau, et. al, "Pseudowire Virtual Circuit Connectivity Verification (VCCV): A Control Channel for Pseudowires ", RFC 5085, December 2007.

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

Authors' Addresses

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

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

George Swallow
Cisco Systems, Inc.
300 Beaver Brook Road
Boxborough , MASSACHUSETTS 01719
United States
Email: swallow@cisco.com

Shaleen Saxena
Cisco Systems, Inc.
1414 Massachusetts Avenue
Boxborough , MASSACHUSETTS 01719
United States
Email: ssaxena@cisco.com

Vishwas Manral
Hewlett Packard Co.
19111 Pruneridge Ave,
Cupertino, CA 95014 USA
United States
EMail: vishwas.manral@hp.com

Michael Wildt
Cisco Systems, Inc.
1414 Massachusetts Avenue
Boxborough , MASSACHUSETTS 01719
United States
Email: mwildt@cisco.com

Sam Aldrin
Huawei Technologies, Inc.
1188 Central Express Way,
Santa Clara, CA 95051
United States
Email: aldrin.ietf@gmail.com

Network Working Group
Internet Draft
Intended status: Standard Track

J.He
Huawei Technologies

H.Li
China Mobile

E. Bellagamba
Ericsson

Expires: March 2012 September 13, 2011

Indication of Client Failure in MPLS-TP
draft-ietf-mpls-tp-csf-02.txt

Status of this Memo

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

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

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

The list of current Internet-Drafts can be accessed at http://www.ietf.org/ietf/1id-abstracts.txt

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

This Internet-Draft will expire on March 13, 2012.

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 in effect on the date of publication of this document (http://trustee.ietf.org/license-info).

Please review these documents carefully, as they describe your rights and restrictions with respect to this document.

Abstract

This document describes a Multi-Protocol Label Switching Transport Profile (MPLS-TP) Operations, Administration and Maintenance (OAM) protocol to propagate a client failure indication across an MPLS-TP network in the case that propagation of failure status in the client layer is not supported as required in [RFC5860].

Table of Contents

1. Introduction

In transport networks, OAM functions are important and fundamental to ease operational complexity, enhance network availability and meet service performance objectives. This is achieved through automatic detection, handling, diagnosis, appropriate reporting of defects and performance monitoring.

As defined in [RFC 5860] MPLS-TP OAM MUST provide a function to enable the propagation, from edge to edge of an MPLS-TP network, of information pertaining to a client (i.e., external to the MPLS-TP network) defect or fault condition detected at an End Point of a PW or LSP, if the client layer OAM functionality does not provide an alarm notification/propagation functionality (e.g. not needed in the

original application of the client signal, or the signal was originally at the bottom of the layer stack and it was not expected to be transported over a server layer), while such an indication is needed by the downstream.

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

According to [RFC 5921], MPLS-TP supports two native service adaptation mechanisms via:

1) a Pseudowire, to emulate certain services, for example, Ethernet, Frame Relay, or PPP / High-Level Data Link Control (HDLC).

2) an LSP, to provide adaptation for any native service traffic type supported by [RFC3031] and [RFC3032]. Examples of such traffic types include IP packets and MPLS-labeled packets (i.e.: PW over LSP, or IP over LSP).

As to the first adaptation mechanism via a PW, the mechanism of CSF function to support propagation of client failure indication follows [static-pw-status]. The PW status relevant to CSF function is AC fault as defined in [RFC 4447] and [RFC 4446].

As to the second adaptation mechanism via LSP, the mechanism is detailed in this draft and is used in case the client of MPLS-TP can not provide itself with such failure notification/propagation.

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].

2.1. Terminology

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

ACH: Associated Channel Header

AIS: Alarm Indication Signal

CSF: Client Signal Fail indication

FDI: Forward Defect Indication

G-ACh: Generic Associated Channel

GAL: G-ACh Label

LSR: Label Switching Router

MEP: Maintenance Entity Group End Point

MIP: Maintenance Entity Group Intermediate Point

OAM: Operations, Administration, and Maintenance

MPLS-TP: MPLS Transport Profile

PW: Pseudowire

RDI: Remote Defect Indication

3. Mechanisms of CSF

3.1. General

Client Signal Fail(CSF) indication provides a function to enable a MEP to propagate a client failure indication to its peer MEP across a MPLS-TP network in case the client service itself does not support propagation of its failure status. A MIP is not intended to generate or process CSF information.

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

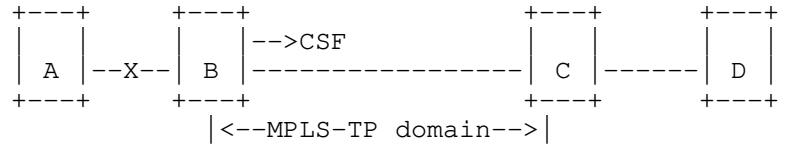


Figure 1 Use case of CSF

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

Client Node D connects to MPLS-TP Node C. Node B and C support MPLS-TP MEP function.

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

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

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

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

3.2. Transmission of CSF

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

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

- Stopping of the transmission of CSF signal but forwarding client data frames, or
- Forwarding CSF PDUs with a clearance indication.

Transmission of packets with CSF information can be enabled or disabled on a MEP (e.g. through management plane).

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

The period of CSF transmission is client and application specific. Examples are as follows:

- 3.33ms: for protection switching application.
- 1s: for fault management application.

However, the value 0 is invalid.

3.3. Reception of CSF

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

CSF condition is cleared when the receiving MEP

- does not receive CSF signal within an interval of N times the CSF transmission period (Suggested value of N is 3.5), or
- receives a valid client data frame, or
- receives CSF PDU with CSF-Clear information

3.4. Configuration of CSF

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

CSF transmission period - this is application dependent. Examples are 3.3 ms and 1s.

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

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

4. Frame format of CSF

Figure 2 depicts the frame format of CSF. CSF PDUs are encapsulated using the ACH, according to [RFC 5586]. GAL is used as an alert based exception mechanism to differentiate CSF packets (with ACH as G-ACh packets) from user-plane packets as defined in [RFC 5586].

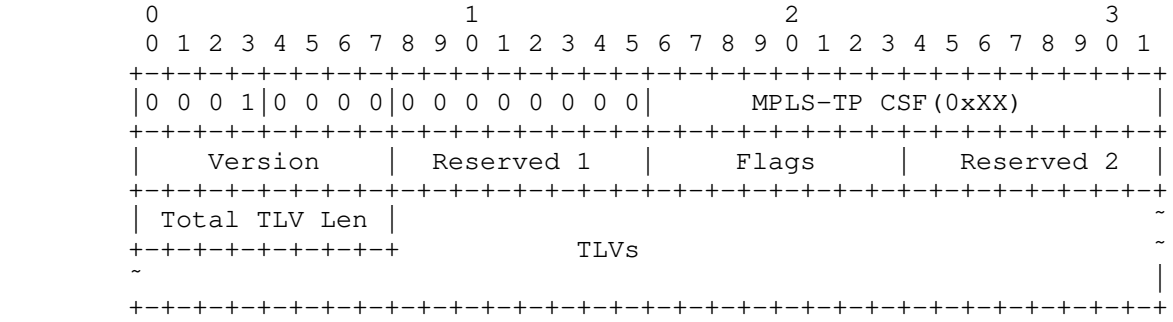


Figure 2 Frame format of CSF

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

- first nibble: set to 0001b to indicate a control channel associated with a PW, a LSP or a Section;

- ACH Version(bits 4 to 7): set to 0, as specified in [RFC 5586]

- ACH Reserved (bits 8 to 15): set to 0 and ignored on reception, as specified in [RFC 5586];

- ACH Channel Type (Bits 16 to 31): value 0xXX identifies the payload as CSF PDU. To be assigned by IANA.

- CSF Version (Bits 32 to 39): Set to 0;

- CSF Reserved 1 (Bits 40 to 47): This field MUST be set to zero on transmission and ignored on receipt;

- CSF Reserved 2 (Bits 56 to 63): This field MUST be set to zero on transmission and ignored on receipt;

- Total TLV Length: Total of all included TLVs. No TLVs are defined currently. The value is 0.

- TLVs: No TLVs are defined currently.

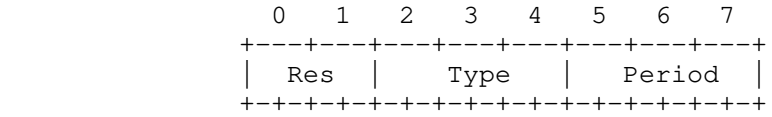


Figure 3 Format of Flags in CSF PDU

Figure 3 depicts the format of Flags in CSF PDU.

- Flag Reserved (Bits 48 to 49): Set to 0;

- Type (Bits 50 to 52): Set to the following values to indicate CSF types

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

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

5. Consequent actions

The primary intention of CSF is to transport a client signal fail condition at the input of the MPLS-TP network to the output port of the MPLS-TP network for clients that do not have alarm notification/propagation mechanism defined.

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

6. Security Considerations

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

7. IANA Considerations

MPLS-TP CSF function requires a new Associated Channel Type to be assigned by IANA from the Pseudowire Associated Channel Types Registry.

Registry:

Value	Description
0xXX	MPLS-TP Client Signal Fail indication (CSF)

8. Acknowledgments

The authors would like to thank Haiyan Zhang, Adrian Farrel, Loa Andersson, Matthew Bocci, Andy Malis and Thomas D. Nadeau for their guidance and input to this work.

9. References

9.1. Normative References

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

[RFC3031] Rosen, E., Viswanathan, A., and R. Callon, "Multiprotocol Label Switching Architecture", RFC 3031, January 2001.

[RFC3032] Rosen, E., Tappan, D., Fedorkow, G., Rekhter, Y., Farinacci, D., Li, T., and A. Conta, "MPLS Label Stack Encoding", RFC 3032, January 2001.

[RFC4446] Martini, L., "IANA Allocations for Pseudowire Edge to Edge Emulation (PWE3)", RFC4446, April 2006

[RFC4447] Martini, L., et al., "Pseudowire Setup and Maintenance Using the Label Distribution Protocol (LDP)", RFC4447, April 2006.

[RFC5586] Vigoureux, M., Bocci, M., Swallow, G., Ward, D., and R. Aggarwal, "MPLS Generic Associated Channel", RFC5586, June 2009

[ITU-T Recommendation G.7041] "Generic framing procedure (GFP)", ITU-T G.7041, October 2008

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

[RFC 5860] Vigoureux, M., Ward, D., and M. Betts, "Requirements for OAM in MPLS Transport Networks", RFC5860, May 2010

[RFC 5921] Bocci, M., Bryant, S., and D. Frost, "A Framework for MPLS in Transport Networks", RFC 5921, July 2010

[static-pw-status] Martini, L., Swallow, G., Heron, G., and M. Bocci, "Pseudowire Status for Static Pseudowires", draft-ietf-pwe3-static-pw-status-06 (work in progress), July 2011

9.2. Informative References

[MPLS-TP OAM Frmk] Busi,I., and D. Allan, "MPLS-TP OAM Framework and Overview", draft-ietf-mpls-tp-oam-framework-11(work in progress), February 2011

[Rosetta stone] Van Helvoort, H., Andersson, L., Sprecher, N., "A Thesaurus for the Terminology used in Multiprotocol Label Switching Transport Profile (MPLS-TP) drafts/RFCs and ITU-T's Transport Network Recommendations", draft-ietf-mpls-tp-rosetta-stone-04 (work in progress), June 2011

10. Authors' Addresses

Jia He
Huawei Technologies Co., Ltd.

Email: hejia@huawei.com

Han Li
China Mobile Communications Corporation

Email: lihan@chinamobile.com

Elisa Bellagamba
Ericsson

Email: elisa.bellagamba@ericsson.com

Intellectual Property

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

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

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

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

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

Disclaimer of Validity

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

Copyright Notice

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

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

MPLS D. Frost, Ed.
Internet-Draft S. Bryant, Ed.
Intended status: Informational Cisco Systems
Expires: January 20, 2012 July 19, 2011

A Packet Loss and Delay Measurement Profile for MPLS-based Transport Networks
draft-ietf-mpls-tp-loss-delay-profile-04

Abstract

Procedures and protocol mechanisms to enable efficient and accurate measurement of packet loss, delay, and throughput in MPLS networks are defined in RFC XXXX.

The MPLS Transport Profile (MPLS-TP) is the set of MPLS protocol functions applicable to the construction and operation of packet-switched transport networks.

This document describes a profile of the general MPLS loss, delay, and throughput measurement techniques that suffices to meet the specific requirements of MPLS-TP.

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

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

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

[RFC Editor, please replace XXXX with the RFC number assigned to draft-ietf-mpls-loss-delay.]

Status of this Memo

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

Internet-Drafts are working documents of the Internet Engineering

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

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

This Internet-Draft will expire on January 20, 2012.

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.

1. Introduction

Procedures for the measurement of packet loss, delay, and throughput in MPLS networks are defined in [I-D.ietf-mpls-loss-delay]. This document describes a profile, i.e. a simplified subset, of these procedures that suffices to meet the specific requirements of MPLS-based transport networks [RFC5921] as defined in [RFC5860]. This profile is presented for the convenience of implementors who are concerned exclusively with the transport network context.

The use of the profile specified in this document is purely optional. Implementors wishing to provide enhanced functionality that is within the scope of [I-D.ietf-mpls-loss-delay] but outside the scope of this profile may do so, whether or not the implementation is restricted to the transport network context.

The assumption of this profile is that the devices involved in a measurement operation are configured for measurement by a means external to the measurement protocols themselves, for example via a Network Management System (NMS) or separate configuration protocol. The manageability considerations in [I-D.ietf-mpls-loss-delay] apply,

and further information on MPLS-TP network management can be found in [RFC5950].

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

2. MPLS-TP Measurement Considerations

The measurement considerations discussed in Section 2.9 of [I-D.ietf-mpls-loss-delay] apply also in the context of MPLS-TP, except for the following, which pertain to topologies excluded from MPLS-TP:

- Equal Cost Multipath considerations (Section 2.9.4 of [I-D.ietf-mpls-loss-delay])
- Considerations for direct Loss Measurement (LM) in the presence of Label Switched Paths constructed via the Label Distribution Protocol (LDP) or utilizing Penultimate Hop Popping (Section 2.9.8 of [I-D.ietf-mpls-loss-delay])

3. Packet Loss Measurement (LM) Profile

When an LM session is externally configured, the values of several protocol parameters can be fixed in advance at the endpoints involved in the session, so that negotiation of these parameters is not required. These parameters, and their default values as specified by this profile, are as follows:

Parameter	Default Value
Query control code	In-band response requested
Byte/packet Count (B) Flag	Packet count
Traffic-Class-specific (T) Flag	Traffic-class-scoped
Origin Timestamp Format (OTF)	Truncated IEEE 1588v2

A simple implementation may assume that external configuration will ensure that both ends of the communication are using the default values for these parameters. Implementations are, however, strongly advised to validate the values of these parameters in received messages so that configuration inconsistencies can be detected and reported.

LM message rates (and test message rates, when inferred LM is used) should be configurable by the network operator on a per-channel basis. The following intervals should be supported:

Message Type	Supported Intervals
LM Message	100 milliseconds, 1 second, 10 seconds, 1 minute, 10 minutes
Test Message	10 milliseconds, 100 milliseconds, 1 second, 10 seconds, 1 minute

4. Packet Delay Measurement (DM) Profile

When a DM session is externally configured, the values of several protocol parameters can be fixed in advance at the endpoints involved in the session, so that negotiation of these parameters is not required. These parameters, and their default values as specified by this profile, are as follows:

Parameter	Default Value
Query control code	In-band response requested
Querier Timestamp Format (QTF)	Truncated IEEE 1588v2
Responder Timestamp Format (RTF)	Truncated IEEE 1588v2
Responder's Preferred Timestamp Format (RPTF)	Truncated IEEE 1588v2

This profile uses the MPLS Delay Measurement (DM) Channel Type in the Associated Channel Header (ACH).

A simple implementation may assume that external configuration will ensure that both ends of the communication are using the default values for these parameters. Implementations are, however, strongly advised to validate the values of these parameters in received messages so that configuration inconsistencies can be detected and reported.

DM message rates should be configurable by the network operator on a per-channel basis. The following message intervals should be supported: 1 second, 10 seconds, 1 minute, 10 minutes.

5. Security Considerations

This document delineates a subset of the procedures specified in [I-D.ietf-mpls-loss-delay], and as such introduces no new security considerations in itself. The security considerations discussed in

[I-D.ietf-mpls-loss-delay] apply also to the profile presented in this document. General considerations for MPLS-TP network security can be found in [I-D.ietf-mpls-tp-security-framework].

6. IANA Considerations

This document introduces no new IANA considerations.

7. References

7.1. Normative References

[I-D.ietf-mpls-loss-delay]
Frost, D. and S. Bryant, "Packet Loss and Delay Measurement for MPLS Networks", draft-ietf-mpls-loss-delay-03 (work in progress), June 2011.

[RFC5586] Bocci, M., Vigoureux, M., and S. Bryant, "MPLS Generic Associated Channel", RFC 5586, June 2009.

[RFC5860] Vigoureux, M., Ward, D., and M. Betts, "Requirements for Operations, Administration, and Maintenance (OAM) in MPLS Transport Networks", RFC 5860, May 2010.

7.2. Informative References

[I-D.ietf-mpls-tp-security-framework]
Fang, L., Niven-Jenkins, B., and S. Mansfield, "MPLS-TP Security Framework", draft-ietf-mpls-tp-security-framework-01 (work in progress), May 2011.

[RFC5921] Bocci, M., Bryant, S., Frost, D., Levrau, L., and L. Berger, "A Framework for MPLS in Transport Networks", RFC 5921, July 2010.

[RFC5950] Mansfield, S., Gray, E., and K. Lam, "Network Management Framework for MPLS-based Transport Networks", RFC 5950, September 2010.

Authors' Addresses

Dan Frost (editor)
Cisco Systems

Email: danfrost@cisco.com

Stewart Bryant (editor)
Cisco Systems

Email: stbryant@cisco.com

Network Working Group
Internet-Draft
Intended status: Informational
Expires: Secptember 13, 2012

D. King (Editor)
Old Dog Consulting
M. Venkatesan (Editor)
Aricent

April 13, 2012

Multiprotocol Label Switching Transport Profile (MPLS-TP) MIB-based Management Overview

draft-ietf-mpls-tp-mib-management-overview-08.txt

Abstract

A range of Management Information Base (MIB) modules has been developed to help model and manage the various aspects of Multiprotocol Label Switching (MPLS) networks. These MIB modules are defined in separate documents that focus on the specific areas of responsibility of the modules that they describe.

The MPLS Transport Profile (MPLS-TP) is a profile of MPLS functionality specific to the construction of packet-switched transport networks.

This document describes the MIB-based architecture for MPLS-TP, and indicates the interrelationships between different existing MIB modules that can be leveraged for MPLS-TP network management and identifies areas where additional MIB modules are required.

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/1id-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 September 13, 2012.

Copyright Notice

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

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

Table of Contents

1. Introduction

The MPLS Transport Profile (MPLS-TP) is a packet transport technology based on a profile of the MPLS functionality specific to the construction of packet-switched transport networks. MPLS is described in [RFC3031] and requirements for MPLS-TP are specified in [RFC5654].

A range of Management Information Base (MIB) modules has been developed to help model and manage the various aspects of Multiprotocol Label Switching (MPLS) networks. These MIB modules

are defined in separate documents that focus on the specific areas of responsibility for the modules that they describe.

An MPLS-TP network can be operated via static provisioning of transport paths, Label Switched Paths (LSPs) and Pseudowires (PW). Or the elective use of a Generalized MPLS (GMPLS) control plane to support dynamic provisioning of transport paths, LSPs and PWs.

This document describes the MIB-based management architecture for MPLS, as extended for MPLS-TP. The document also indicates the interrelationships between existing MIB modules that should be leveraged for MPLS-TP network management and identifies areas where additional MIB modules are required.

Note that [RFC5951] does not specify a preferred management interface protocol to be used as the standard protocol for managing MPLS-TP networks.

1.1 MPLS-TP Management Function

The management of the MPLS-TP networks is separable from that of its client networks so that the same means of management can be used regardless of the client. The management function of MPLS-TP includes fault management, configuration management, performance monitoring, and security management.

The purpose of the management function is to provide control and monitoring of the MPLS transport profile protocol mechanisms and procedures. The requirements for the network management functionality are found in [RFC5951]. A description of the network and element management architectures that can be applied to the management of MPLS-based transport networks is found in [RFC5950].

2. Terminology

This document also uses terminology from the MPLS architecture document [RFC3031], PWE3 architecture [RFC4805], and the following MPLS related MIB modules: MPLS TC MIB [RFC3811], MPLS LSR MIB [RFC3813], MPLS TE MIB [RFC3812], MPLS LDP MIB [RFC3815], MPLS FTN MIB [RFC3814] and TE LINK MIB [RFC4220].

3. The SNMP Management Framework

Managed objects are accessed via a virtual information store, termed the Management Information Base or MIB. MIB objects are generally accessed through the Simple Network Management Protocol (SNMP).

Objects in the MIB are defined using the mechanisms defined in the Structure of Management Information (SMI).

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

This document discusses MIB modules that are compliant to the SMIv2, which is described in [RFC2578], [RFC2579] and [RFC2580].

4. Overview of Existing Work

This section describes the existing tools and techniques for managing and modeling MPLS networks, devices, and protocols. It is intended to provide a description of the tool kit that is already available.

Section 5 of this document outlines the applicability of existing MPLS MIB modules to MPLS-TP, describes the optional use of GMPLS MIB modules in MPLS-TP networks, and examines the additional MIB modules and objects that would be required for managing an MPLS-TP network.

4.1. MPLS Management Overview and Requirements

[RFC4378] outlines how data plane protocols can assist in providing the Operations and Management (OAM) requirements outlined in [RFC4377] and how it is applied to the management functions of fault, configuration, accounting, performance, and security (commonly known as FCAPS) for MPLS networks.

[RFC4221] describes the management architecture for MPLS. In particular, it describes how the managed objects defined in various MPLS-related MIB modules model different aspects of MPLS, as well as the interactions and dependencies between each of these MIB modules.

[RFC4377] describes the requirements for user and data plane OAM and applications for MPLS.

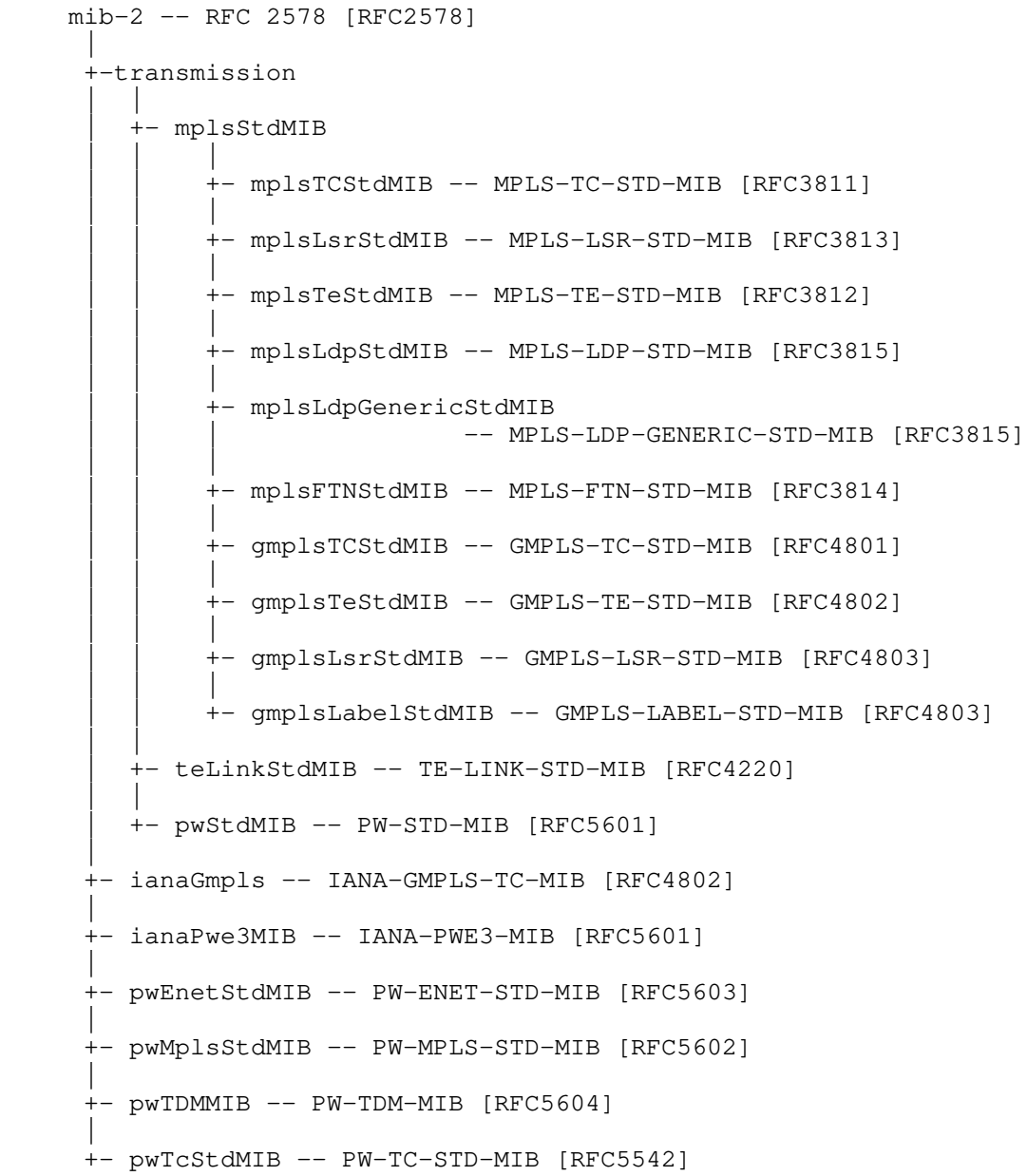
[RFC5654] describes the requirements for the optional use of a control plane to support dynamic provisioning of MPLS-TP transport paths. The MPLS-TP LSP control plane is based on GMPLS and is described in [RFC3945].

4.2. An Introduction to the MPLS and Pseudowire MIB Modules

4.2.1. Structure of the MPLS MIB OID Tree

The MPLS MIB Object Identifiers (OID) tree has the following

structure. It is based on the tree originally set out in section 4.1 of [RFC4221] and has been enhanced to include other relevant MIB modules.



Note: The OIDs for MIB modules are assigned and managed by IANA. They can be found in the referenced MIB documents.

4.2.2. Textual Convention Modules

MPLS-TC-STD-MIB [RFC3811], GMPLS-TC-STD-MIB [RFC4801], IANA-GMPLS-TC-MIB [RFC4802] and PW-TC-STD-MIB [RFC5542] contains the Textual Conventions for MPLS and GMPLS networks. These Textual Conventions should be imported by MIB modules which manage MPLS and GMPLS networks. Section 4.2.11. highlights dependencies on additional external MIB modules

4.2.3. Label Switched Path (LSP) Modules

An LSP is a path over which a labeled packet travels across the sequence of LSRs for a given Forward Equivalence Class (FEC). When a packet, with or without label, arrives at an ingress LER of an LSP, it is encapsulated with the label corresponding to the FEC and sent across the LSP. The labeled packet traverses across the LSRs and arrives at the egress LER of the LSP, where, it gets forwarded depending on the packet type it came with. LSPs could be nested using label stacking, such that, an LSP could traverse over another LSP. A further description of an LSP can be found in [RFC3031].

MPLS-LSR-STD-MIB [RFC3813] describes the required objects to define the LSP.

4.2.4. Label Edge Router (LER) Modules

Ingress and Egress LSRs of an LSP are known as Label Edge Routers (LER). An ingress LER takes the incoming unlabeled or labeled packets and encapsulates it with the corresponding label of the LSP it represents, and forwards it, over to the adjacent LSR of the LSP. Each FEC is mapped to a label forwarding entry, so that packet could be encapsulated with one or more label entries, referred as label stack.

The packet traverses across the LSP, and upon reaching the Egress LER, further action will be taken to handle the packet, depending on the packet it received. MPLS Architecture [RFC3031] details the functionality of an Ingress and Egress LERs.

MPLS-FTN-STD-MIB [RFC3814] describes the managed objects for mapping FEC to label bindings.

4.2.5. Label Switching Router (LSR) Modules

A router which performs MPLS forwarding is known as an LSR. An LSR receives a labelled packet and performs forwarding action based on the label received.

LSR maintains a mapping of an incoming label and incoming interface

to one or more outgoing label and outgoing interfaces in its forwarding database. When a labelled packet is received, LSR examines the topmost label in the label stack and then does 'swap', 'push' or 'pop' operation based on the contents.

MPLS-LSR-STD-MIB [RFC3813] describes the managed objects for modeling a Multiprotocol Label Switching (MPLS) [RFC3031] LSR. MPLS-LSR-STD-MIB [RFC3813] contains the managed objects to maintain mapping of in-segments to out-segments.

4.2.6. Pseudowire Modules

The PW (Pseudowire) MIB architecture provides a layered modular model into which any supported emulated service such as Frame Relay, ATM, Ethernet, TDM and SONET/SDH can be connected to any supported Packet Switched Network (PSN) type. This MIB architecture is modeled based on PW3 architecture [RFC3985].

Emulated Service Layer, Generic PW Layer and PSN VC Layer constitute the different layers of the model. A combination of the MIB modules belonging to each layer provides the glue for mapping the emulated service onto the native PSN service. At least three MIB modules each belonging to a different layer are required to define a PW emulated service.

- Service-Specific module is dependent on the emulated signal type and helps in modeling emulated service layer.

PW-ENET-STD-MIB [RFC5603] describes a model for managing Ethernet pseudowire services for transmission over a PSN. This MIB module is generic and common to all types of PSNs supported in the Pseudowire Emulation Edge-to-Edge (PWE3) Architecture [RFC3985], which describes the transport and encapsulation of L1 and L2 services over supported PSN types.

In particular, the MIB module associates a port or specific VLANs on top of a physical Ethernet port or a virtual Ethernet interface (for Virtual Private LAN Service (VPLS)) to a point-to-point PW. It is complementary to the PW-STD-MIB [RFC5601], which manages the generic PW parameters common to all services, including all supported PSN types.

PW-TDM-MIB [RFC5604] describes a model for managing TDM pseudowires, i.e., TDM data encapsulated for transmission over a Packet Switched Network (PSN). The term TDM in this document is limited to the scope of Plesiochronous Digital Hierarchy (PDH). It is currently specified to carry any TDM Signals in either Structure Agnostic Transport mode (E1, T1, E3, and T3) or in Structure Aware Transport mode (E1, T1, and NxDS0) as defined in the Pseudowire

Emulation Edge-to-Edge (PWE3) TDM Requirements document [RFC4197].

- Generic PW Module configures general parameters of the PW that are common to all types of emulated services and PSN types.

PW-STD-MIB [RFC5601] defines a MIB module that can be used to manage pseudowire (PW) services for transmission over a Packet Switched Network (PSN) [RFC3931] [RFC4447]. This MIB module provides generic management of PWs that is common to all types of PSN and PW services defined by the IETF PWE3 Working Group.

- PSN-specific module associate the PW with one or more "tunnels" that carry the service over the PSN. There is a different module for each type of PSN.

PW-MPLS-STD-MIB [RFC5602] describes a model for managing pseudowire services for transmission over different flavors of MPLS tunnels. The general PW MIB module [RFC5601] defines the parameters global to the PW regardless of the underlying Packet Switched Network (PSN) and emulated service. This document is applicable for PWs that use MPLS PSN type in the PW-STD-MIB. Additionally this document describes the MIB objects that define pseudowire association to the MPLS PSN, that is not specific to the carried service.

Together, [RFC3811], [RFC3812] and [RFC3813] describe the modeling of an MPLS tunnel, and a tunnel's underlying cross-connects. This MIB module supports MPLS-TE PSN, non-TE MPLS PSN (an outer tunnel created by the Label Distribution Protocol (LDP) or manually), and MPLS PW label only (no outer tunnel).

4.2.7. Routing and Traffic Engineering

In MPLS traffic engineering, it's possible to specify explicit routes or choose routes based on QOS metrics in setting up a path such that some specific data can be routed around network hot spots. TE LSPs can be setup through a management plane or a control plane.

MPLS-TE-STD-MIB [RFC3812] describes managed objects for modeling a Multiprotocol Label Switching (MPLS) [RFC3031] based traffic engineering. This MIB module should be used in conjunction with the companion document [RFC3813] for MPLS based traffic engineering configuration and management.

4.2.8. Resiliency

The purpose of MPLS resiliency is to ensure minimal interruption to traffic when the failure occurs within the system or network.

Various components of MPLS resiliency solutions are;

1) Graceful restart in LDP and RSVP-TE modules,
2) Make Before Break,
3) Protection Switching for LSPs,
4) Fast ReRoute for LSPs,
5) PW redundancy.

The MIB modules below only support MIB based management for MPLS resiliency.

MPLS Fast Reroute (FRR) is a restoration network resiliency mechanism used in MPLS TE to redirect the traffic onto the backup LSP's in 10s of milliseconds in case of link or node failure across the LSP.

MPLS-FRR-GENERAL-STD-MIB [draft-ietf-mpls-fastreroute-mib-14] contains objects that apply to any MPLS LSR implementing MPLS TE fast reroute functionality.

MPLS-FRR-ONE2ONE-STD-MIB [draft-ietf-mpls-fastreroute-mib-14] contains objects that apply to one-to-one backup method.
MPLS-FRR-FACILITY-STD-MIB [draft-ietf-mpls-fastreroute-mib-14] contains objects that apply to facility backup method.

Protection Switching mechanisms have been designed to provide network resiliency for MPLS network. Different types of protection switching mechanisms such as 1:1, 1:N, 1+1 have been designed.

4.2.9. Fault Management and Performance Management

MPLS manages the LSP and pseudowire faults through the use of LSP ping [RFC4379], VCCV [RFC5085], BFD for LSPs [RFC5884] and BFD for VCCV [RFC5885] tools.

Current MPLS focuses on the in and/or out packet counters, errored packets, discontinuity time.

Some of the MPLS and Pseudowire performance tables used for performance management are given below.

mplsTunnelPerfTable [RFC3812] provides several counters (packets forwarded, packets dropped because of errors) to measure the performance of the MPLS tunnels.

mplsInterfacePerfTable [RFC3813] provides performance information (incoming and outgoing labels in use and lookup failures) on a per-interface basis.

mplsInSegmentPerfTable [RFC3813] contains statistical information (total packets received by the insegment, total errored packets received, total packets discarded, discontinuity time) for incoming

MPLS segments to an LSR.

mplsOutSegmentPerfTable [RFC3813] contains statistical information (total packets received, total errored packets received, total packets discarded, discontinuity time) for outgoing MPLS segments from an LSR.

mplsFTNPerfTable [RFC3814] contains performance information for the specified interface and an FTN entry mapped to this interface.

mplsLdpEntityStatsTable [RFC3815] and mplsLdpSessionStatsTable [RFC3815] contain statistical information (session attempts, errored packets, notifications) about an LDP entity.

pwPerfCurrentTable [RFC5601], pwPerfIntervalTable [RFC5601], pwPerf1DayIntervalTable [RFC5601] provides pseudowire performance information (in and/or out packets) based on time (current interval, preconfigured specific interval, 1day interval).

pwEnetStatsTable [RFC5603] contains statistical counters specific for Ethernet PW.

pwTDMPerfCurrentTable [RFC5604], pwTDMPerfIntervalTable [RFC5604] and pwTDMPerf1DayIntervalTable [RFC5604] contain statistical informations accumulated per 15-minute, 24 hour, 1 day respectively.

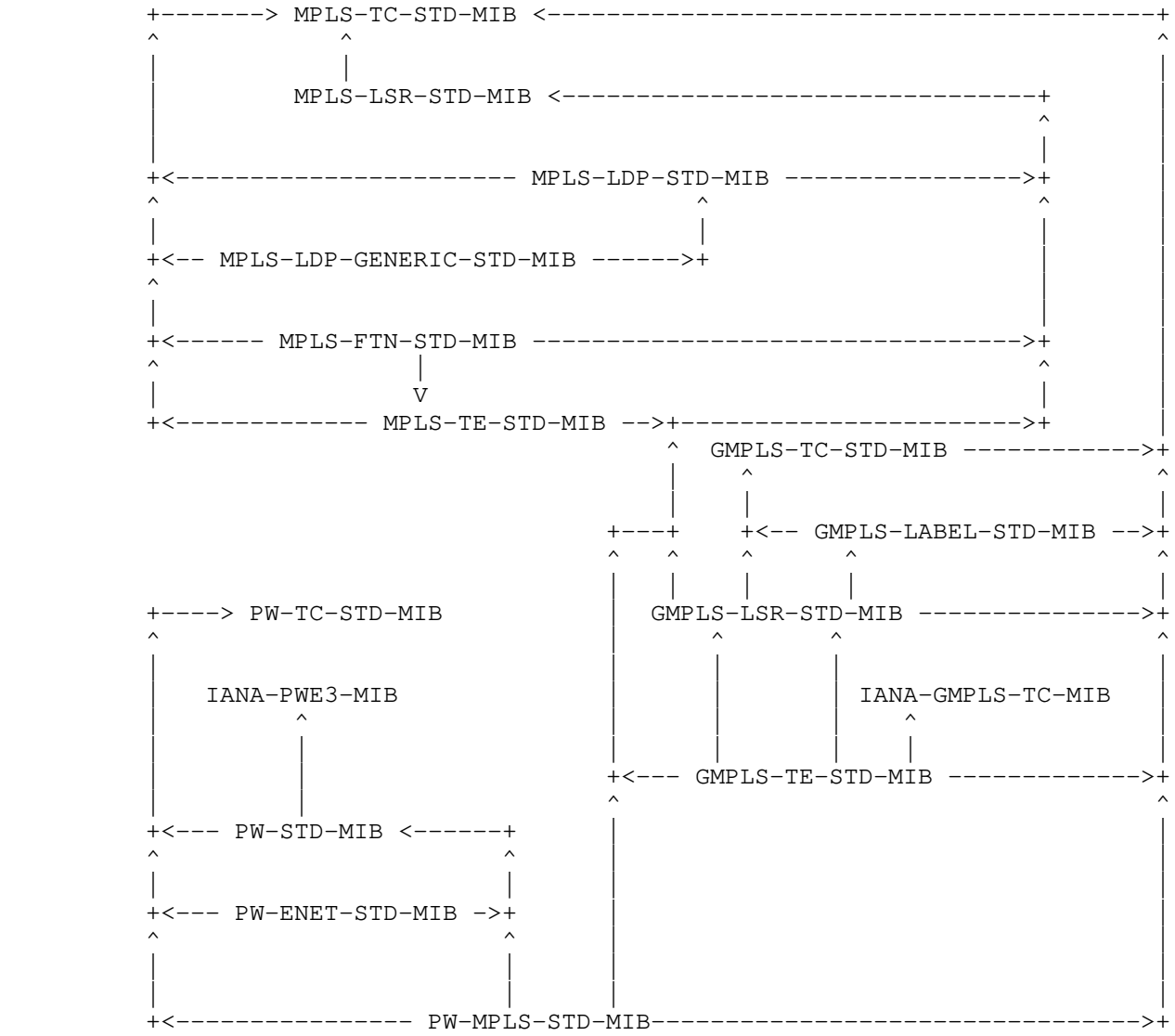
gmplsTunnelErrorTable [RFC4802] and gmplsTunnelReversePerfTable [RFC4802] provides information about performance errored packets and in/out packet counters.

4.2.10. MIB Module Interdependencies

This section provides an overview of the relationship between the MPLS MIB modules for managing MPLS networks. More details of these relationships are given below.

[RFC4221] mainly focuses on the MPLS MIB module interdependencies, this section also highlights the GMPLS and PW MIB modules interdependencies.

The relationship "A --> B" means A depends on B and that MIB module A uses an object, object identifier, or textual convention defined in MIB module B, or that MIB module A contains a pointer (index or RowPointer) to an object in MIB module B.



Thus:

- All the MPLS MIB modules depend on MPLS-TC-STD-MIB.

- All the GMPLS MIB modules depend on GMPLS-TC-STD-MIB.

- All the PW MIB modules depend on PW-TC-STD-MIB.

- MPLS-LDP-STD-MIB, MPLS-TE-STD-MIB, MPLS-FTN-STD-MIB, GMPLS-LSR-STD-MIB, and PW-MPLS-STD-MIB contain references to objects in MPLS-LSR-STD-MIB.

- MPLS-LDP-GENERIC-STD-MIB contains references to objects in MPLS-LDP-STD-MIB.

- MPLS-FTN-STD-MIB, PW-MPLS-STD-MIB, and GMPLS-TE-STD-MIB contain references to objects in MPLS-TE-STD-MIB.

- PW-MPLS-STD-MIB, and PW-ENET-STD-MIB contains references to objects in PW-STD-MIB.

- PW-STD-MIB contains references to objects in IANA-PWE3-MIB.

- GMPLS-TE-STD-MIB contains references to objects in IANA-GMPLS-TC-MIB.

- GMPLS-LSR-STD-MIB contains references to objects in GMPLS-LABEL-STD-MIB.

Note that there is a textual convention (MplsIndexType) defined in MPLS-LSR-STD-MIB that is imported by MPLS-LDP-STD-MIB.

4.2.11. Dependencies on External MIB Modules

With the exception of MPLS-TC-STD-MIB, all the MPLS MIB modules have dependencies on the Interfaces MIB [RFC2863]. MPLS-FTN-STD-MIB references IP-capable interfaces on which received traffic is to be classified using indexes in the Interface Table (ifTable) of IF-MIB [RFC2863]. The other MPLS MIB modules reference MPLS-capable interfaces in ifTable.

The Interfaces Group of IF-MIB [RFC2863] defines generic managed objects for managing interfaces. The MPLS MIB modules contain media-specific extensions to the Interfaces Group for managing MPLS interfaces.

The MPLS MIB modules assume the interpretation of the Interfaces Group to be in accordance with [RFC2863], which states that ifTable contains information on the managed resource's interfaces and that each sub-layer below the internetwork layer of a network interface is considered an interface. Thus, the MPLS interface is represented as an entry in ifTable.

The interrelation of entries in ifTable is defined by the Interfaces Stack Group defined in [RFC2863].

The MPLS MIB modules have dependencies with the TE-LINK-STD-MIB for maintaining the traffic engineering information.

The MPLS MIB modules depend on the constrained shortest path first (CSPF) module to obtain the path required for an MPLS tunnel to reach the end point of the tunnel and Bidirectional Forwarding Detection (BFD) module to verify the data-plane failures of LSPs and PWs.

Finally, all of the MIB modules import standard textual conventions such as integers, strings, timestamps, etc., from the MIB modules in which they are defined.

5. Applicability of MPLS MIB modules to MPLS-TP

This section highlights gaps in existing MPLS MIB modules in order to determine extensions or additional MIB modules that are required to support MPLS-TP in MPLS networks

[RFC5951] specifies the requirements for the management of equipment used in networks supporting an MPLS-TP. It also details the essential network management capabilities for operating networks consisting of MPLS-TP equipment.

[RFC5950] provides the network management framework for MPLS-TP. The document explains how network elements and networks that support MPLS-TP can be managed using solutions that satisfy the requirements defined in [RFC5951]. The relationship between MPLS-TP management and OAM is described in the MPLS-TP framework [RFC5950] document.

The MPLS MIB modules MPLS-TE-STD-MIB [RFC3812], PW-STD-MIB [RFC5601] and MPLS-LSR-STD-MIB [RFC3813] and their associated MIB modules are reused for MPLS based transport network management.

Fault management and performance management form key parts of the Operations, Administration, and Maintenance (OAM) function. MPLS-TP OAM is described in [MPLS-TP-OAM-FWK].

5.1 MPLS-TP Tunnel

5.1.1 Gap Analysis

MPLS-TP tunnel can be operated over IP and/or ITU-T Carrier Code (ICC) environments, below points capture the gaps in existing MPLS MIB modules for managing the MPLS-TP networks.

- IP based environment
  i. MPLS-TE-STD-MIB [RFC3812] does not support tunnel Ingress/Egress identifier based on Global_ID and Node_ID [RFC6370].
  ii. MPLS-TE-STD-MIB [RFC3812] does not support co-routed/associated bidirectional tunnel configurations.

- ICC based environment
  i. MPLS-TE-STD-MIB [RFC3812] does not support tunnel LSR identifier based on ICC.

5.1.2 Recommendations

- New MIB definitions may be created for Global_Node_ID and/or ICC configurations.

- MPLS-LSR-STD-MIB [RFC3813] MIB modules may be enhanced to identify the nexthop based on MAC address for IP-less environments. OutSegment may be extended to hold the MAC-address also for IP-less environments.

- MPLS-TE-STD-MIB [RFC3812] and MPLS-LSR-STD-MIB may be enhanced to provide static and signalling MIB module extensions for co-routed/associated bidirectional LSPs.

5.2 MPLS-TP Pseudowire

5.2.1 Gap Analysis

MPLS-TP Pseudowire can be operated over IP and/or ICC environments, below points capture the gaps in existing PW MIB modules for managing the MPLS-TP networks.

[RFC6370] specifies an initial set of identifiers to be used in MPLS-TP. These identifiers were chosen to be compatible with existing MPLS, GMPLS, and PW definitions.

- IP based environment
  i. PW-STD-MIB [RFC5601] does not support PW end point identifier based on Global_ID and Node_ID.
  ii. PW-MPLS-STD-MIB [RFC5602] does not support its operation over co-routed/associated bidirectional tunnels.

- ICC based environment
  i. PW-STD-MIB [RFC5601] does not support PW end point identifier based on ICC.

5.2.2 Recommendations

- PW-MPLS-STD-MIB [RFC5602] can be enhanced to operate over co-routed/associated bi-directional tunnel.

5.3 MPLS-TP Sections

5.3.1 Gap Analysis

The existing MPLS MIB modules do not support MPLS-TP sections.

5.3.2 Recommendations

Link specific and/or path/segment specific sections can be supported by enhancing the IF-MIB [RFC2863], MPLS-TE-STD-MIB [RFC3812] and PW-STD-MIB [RFC5601] MIB modules.

5.4 MPLS-TP OAM

5.4.1 Gap Analysis

MPLS manages the LSP and pseudowire faults through LSP ping [RFC4379], VCCV [RFC5085], BFD for LSPs [RFC5884] and BFD for VCCV [RFC5885] tools.

The MPLS MIB modules do not support the below MPLS-TP OAM functions:

- o Continuity Check and Connectivity Verification
- o Remote Defect Indication
- o Alarm Reporting
- o Lock Reporting
- o Lock Instruct
- o Client Failure Indication
- o Packet Loss Measurement
- o Packet Delay Measurement

5.4.2 Recommendations

New MIB module for BFD can be created to address all the gaps mentioned in Section 5.4.1. (Gap Analysis).

5.5 MPLS-TP Protection Switching and Recovery

5.5.1 Gap Analysis

An important aspect that MPLS-TP technology provides is protection switching. In general, the mechanism of protection switching can be described as the substitution of a protection or standby facility for a working or primary facility.

The MPLS MIB modules do not provide support for protection switching and recovery of three different topologies (linear, ring and mesh) available.

5.5.2 Recommendations

New MIB modules can be created to address all the gaps mentioned in the 5.5.1 Gap Analysis section.

5.6 MPLS-TP Interfaces

5.6.1 Gap Analysis

As per [RFC6370], an LSR requires identification of the node itself and of its interfaces. An interface is the attachment point to a server layer MPLS-TP section or MPLS-TP tunnel.

The MPLS MIB modules do not provide support for configuring the interfaces within the context of an operator.

5.6.2 Recommendations

New MIB definitions can be created to address the gaps mentioned in the 5.6.1 Gap Analysis section.

6. An Introduction to the MPLS-TP MIB Modules

This section highlights new MIB modules that have been identified as being required for MPLS-TP. This section also provides an overview the purpose of each of the MIB modules within the MIB documents, what it can be used for, and how it relates to the other MIB modules.

Note that each new MIB module (apart from Textual Conventions modules) will contain one or more Compliance Statements to indicate which objects must be supported in what manner to claim a specific level of compliance. Additional text, either in the documents that define the MIB modules or in separate Applicability Statements, will define which Compliance Statements need to be conformed to in order to provide specific MPLS-TP function. This document does not set any requirements in that respect although some recommendations are included in the sections that follow.

6.1 MPLS-TP MIB Modules

6.1.1 NEW MIB Modules for MPLS-TP

Four new MIB modules are identified as follows:

- Textual Conventions for MPLS-TP

- Identifiers for MPLS-TP

- LSR MIB Extensions for MPLS-TP

- TE MIB Extensions for MPLS-TP

Note that the MIB modules mentioned here are applicable for MPLS operations as well.

6.1.2 Textual Conventions for MPLS-TP

A new MIB module needs to be written that will define textual conventions [RFC2579] for MPLS-TP related MIB modules. These conventions allow multiple MIB modules to use the same syntax and format for a concept that is shared between the MIB modules.

For example, MEP identifier is used to identify maintenance entity group end point within MPLS-TP networks. The textual convention representing the MEP identifier should be defined in a new textual convention MIB module.

All new extensions related to MPLS-TP are defined in the MIB module and will be referenced by other MIB modules to support MPLS-TP.

6.1.3 Identifiers for MPLS-TP

New Identifiers describe managed objects that are used to model common MPLS-TP identifiers [RFC6370].

6.1.4 LSR MIB Extensions for MPLS-TP

MPLS-LSR-STD-MIB describes managed objects for modeling an MPLS Label Switching Router (LSR). This puts it at the heart of the management architecture for MPLS.

In the case of MPLS-TP, the MPLS-LSR-STD-MIB is extended to support the MPLS-TP LSP's, which are co-routed or associated bidirectional. This extended MIB is also applicable for modeling MPLS-TP tunnels.

6.1.5 Tunnel Extensions for MPLS-TP

MPLS-TE-STD-MIB describes managed objects that are used to model and manage MPLS Traffic Engineered (TE) Tunnels.

MPLS-TP tunnels are very similar to MPLS-TE tunnels, but are co-routed or associated bidirectionally.

The MPLS-TE-STD-MIB must be extended to support the MPLS-TP specific attributes for the tunnel.

6.2 PWE3 MIB Modules for MPLS-TP

This section provides an overview of Pseudowire extension MIB modules to meet the MPLS based transport network requirements.

6.2.1 New MIB Modules for MPLS-TP Pseudowires

Three new MIB modules are identified as follows:

- Pseudowire Extensions for MPLS-TP

- Pseudowire MPLS Extensions for MPLS-TP

- Pseudowire Textual Conventions for MPLS-TP

6.2.2 Pseudowire Textual Conventions for MPLS-TP

PW-TC-STD-MIB MIB defines textual conventions used for pseudowire (PW) technology and for Pseudowire Edge-to-Edge Emulation (PWE3) MIB Modules. A new textual convention MIB module is required to define textual definitions for MPLS-TP specific Pseudowire attributes.

6.2.3 Pseudowire Extensions for MPLS-TP

PW-STD-MIB describes managed objects for modeling of Pseudowire Edge-to-Edge services carried over a general Packet Switched Network. This MIB module is extended to support MPLS-TP specific attributes related to Pseudowires.

6.2.4 Pseudowire MPLS Extensions for MPLS-TP

PW-MPLS-STD-MIB defines the managed objects for Pseudowire operations over MPLS LSR's. This MIB supports both, manual and dynamically signaled PW's, point-to-point connections, enables the use of any emulated service, MPLS-TE as outer tunnel and no outer tunnel as MPLS-TE.

The newly extended MIB defines the managed objects, extending PW-MPLS-STD-MIB, by supporting with or without MPLS-TP as outer tunnel.

6.3 OAM MIB Modules for MPLS-TP

This section provides an overview of Operations, Administration, and Maintenance (OAM) MIB modules for MPLS LSPs and Pseudowires.

6.3.1 New MIB Modules for OAM for MPLS-TP

Two new MIB modules are identified as follows:

- BFD MIB module

- OAM MIB module

6.3.2 BFD MIB module

BFD-STD-MIB defines managed objects for performing BFD operation in IP networks. This MIB is modeled to support BFD protocol [RFC5880].

A new MIB module needs to be written that will be an extension to BFD-STD-MIB managed objects to support BFD operations on MPLS LSPs and PWs.

6.3.3 Common OAM MIB modules

A new MIB module needs to be written that will define managed objects for OAM maintenance identifiers i.e. Maintenance Entity Group Identifiers (MEG), Maintenance Entity Group End-point (MEP), Maintenance Entity Group Intermediate Point (MIP). Maintenance points are uniquely associated with a MEG. Within the context of a MEG, MEPs and MIPs must be uniquely identified.

6.4. Protection Switching and Recovery MIB Modules for MPLS-TP

This section provides an overview of protection switching and recovery MIB modules for MPLS LSPs and Pseudowires.

6.4.1 New MIB Modules for MPLS Protection Switching and Recovery

Three new MIB modules are identified as follows:

- Linear Protection Switching MIB module

- Ring Protection Switching MIB module

- Mesh Protection Switching MIB module

6.4.2 Linear Protection Switching MIB module

A new MIB module needs to be written that will define managed objects for linear protection switching of MPLS LSPs and Pseudowires.

6.4.3 Ring Protection Switching MIB module

A new MIB module will define managed objects for ring protection switching of MPLS LSPs and Pseudowires.

6.4.4 Mesh Protection Switching MIB module

A new MIB module needs to be written that will define managed objects for Mesh protection switching of MPLS LSPs and Pseudowires.

7. Management Options

This document applies only to scenarios where MIB modules are used to manage the MPLS-TP network. It is not the intention of this document to provide instructions or advice to implementers of management

systems, management agents, or managed entities. It is, however, useful to make some observations about how the MIB modules described above might be used to manage MPLS systems, if SNMP is used in the management interface.

For MPLS specific management options, refer to [RFC4221] Section 12. (Management Options).

8. Security Considerations

This document describes the interrelationships amongst the different MIB modules relevant to MPLS-TP management and as such does not have any security implications in and of itself.

Each IETF MIB document that specifies MIB objects for MPLS-TP must provide a proper security considerations section that explains the security aspects of those objects.

The attention of readers is particularly drawn to the security implications of making MIB objects available for create or write access through an access protocol such as SNMP. SNMPv1 by itself is an insecure environment. Even if the network itself is made secure (for example, by using IPSec), there is no control over who on the secure network is allowed to access the objects in this MIB. It is recommended that the implementers consider the security features as provided by the SNMPv3 framework. Specifically, the use of the User-based Security Model STD 62, RFC3414 [RFC3414], and the View-based Access Control Model STD 62, RFC 3415 [RFC3415], is recommended.

It is then a customer/user responsibility to ensure that the SNMP entity giving access to an instance of each MIB module is properly configured to give access to only those objects, and to those principals (users) that have legitimate rights to access them.

9. IANA Considerations

This document has identified areas where additional MIB modules are neccessary for MPLS-TP. The new MIB modules recommended by this document will require OID assignments from IANA. However, this document makes no specific request for IANA action.

10. Acknowledgements

The authors would like to thank Eric Gray, Thomas Nadeau, Benjamin Niven-Jenkins, Saravanan Narasimhan, Joel Halpern, David Harrington, and Stephen Farrell for their valuable comments.

This document also benefited from review by participants in ITU-T Study Group 15.

11. References

11.1 Normative References

[RFC2863] McCloghrie, K. and F. Kastenholz, "The Interfaces Group MIB using SMIv2", RFC 2863, June 2000.

[RFC3811] Nadeau, T. and J. Cucchiara, "Definition of Textual Conventions and for Multiprotocol Label Switching (MPLS) Management", RFC 3811, June 2004.

[RFC3812] Srinivasan, C., Viswanathan, A., and T. Nadeau, "Multiprotocol Label Switching (MPLS) Traffic Engineering (TE) Management Information Base (MIB)", RFC 3812, June 2004.

[RFC3813] Srinivasan, C., Viswanathan, A., and T. Nadeau, "Multiprotocol Label Switching (MPLS) Label Switching (LSR) Router Management Information Base (MIB)", RFC 3813, June 2004.

[RFC3814] Nadeau, T., Srinivasan, C., and A. Viswanathan, "Multiprotocol Label Switching (MPLS) FEC-To-NHLFE (FTN) Management Information Base", RFC3814, June 2004.

[RFC3815] Cucchiara, J., Sjostrand, H., and Luciani, J., "Definitions of Managed Objects for the Multiprotocol Label Switching (MPLS), Label Distribution Protocol (LDP)", RFC 3815, June 2004.

[RFC4220] Dubuc, M., Nadeau, T., and J. Lang, "Traffic Engineering Link Management Information Base", RFC 4220, November 2005.

[RFC4221] Nadeau, T., Srinivasan, C., and A. Farrel, "Multiprotocol Label Switching (MPLS) Management Overview", RFC 4221, November 2005.

[RFC4801] T. Nadeau and A. Farrel, Ed., "Definitions of Textual Conventions for Generalized Multiprotocol Label Switching (GMPLS) Management", RFC4801, Feb. 2007.

[RFC4802] T. D. Nadeau and A. Farrel, "Generalized Multiprotocol Label Switching (GMPLS) Traffic Engineering Management Information Base", RFC4802, Feb., 2007.

[RFC4803] T. D. Nadeau and A. Farrel, "Generalized Multiprotocol Label Switching (GMPLS) Label Switching Router (LSR) Management Information Base", RFC4803, Feb., 2007.

[RFC5542] Nadeau, T., Ed., Zelig, D., Ed., and O. Nicklass, Ed., "Definitions of Textual Conventions for Pseudowire (PW) Management", RFC 5542, May 2009.

[RFC5601] Nadeau, T., Ed. and D. Zelig, Ed. "Pseudowire (PW) Management Information Base (MIB)", RFC 5601, July 2009.

[RFC5602] Zelig, D., Ed., and T. Nadeau, Ed., "Pseudowire (PW) over MPLS PSN Management Information Base (MIB)", RFC 5602, July 2009.

[RFC5603] Zelig, D., Ed., and T. Nadeau, Ed., "Ethernet Pseudowire (PW) Management Information Base (MIB)", RFC 5603, July 2009.

[RFC5604] Nicklass, O., "Managed Objects for Time Division Multiplexing (TDM) over Packet Switched Networks (PSNs)", RFC5604, July 2009.

11.2 Informative References

[RFC2578] McCloghrie, K., Perkins, D., and J. Schoenwaelder, "Structure of Management Information Version 2 (SMIv2)", STD 58, RFC 2578, April 1999.

[RFC2579] McCloghrie, K., Perkins, D., and J. Schoenwaelder, "Textual Conventions for SMIv2", STD 58, RFC 2579, April 1999.

[RFC2580] McCloghrie, K., Perkins, D., and J. Schoenwaelder, "Conformance Statements for SMIv2", STD 58, RFC 2580, April 1999.

[RFC3031] Rosen, E., Viswanathan, A., and R. Callon, "Multiprotocol Label Switching Architecture", RFC 3031, March 2001.

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

[RFC3414] Blumenthal, U. and B. Wijnen, "User-based Security Model (USM) for version 3 of the Simple Network Management Protocol (SNMPv3)", STD 62, RFC 3414, December 2002.

[RFC3415] Wijnen, B., Presuhn, R., and K. McCloghrie, "View-based Access Control Model (VACM) for the Simple Network Management Protocol (SNMP)", STD 62, RFC 3415, December 2002.

[RFC3812] Srinivasan, C., Viswanathan, A., and T. Nadeau, "Multiprotocol Label Switching (MPLS) Traffic Engineering (TE) Management Information Base (MIB)", RFC 3812, June 2004.

[RFC3813] Srinivasan, C., Viswanathan, A., and T. Nadeau, "Multiprotocol Label Switching (MPLS) Label Switching Router (LSR) Management Information Base (MIB)", RFC 3813, June 2004.

[RFC3931] Lau, J., Townsley, M., and I. Goyret, "Layer Two Tunneling Protocol - Version 3 (L2TPv3)", RFC 3931, March 2005.

[RFC3945] Mannie, E. et.al., "Generalized Multi-Protocol Label Switching (GMPLS) Architecture", IETF RFC 3945, October 2004.

[RFC3985] Bryant, S. and P. Pate, "Pseudo Wire Emulation Edge-to-Edge (PWE3) Architecture", RFC 3985, March 2005.

[RFC4197] Riegel, M., "Requirements for Edge-to-Edge Emulation of Time Division Multiplexed (TDM) Circuits over Packet Switching Networks", RFC4197, October 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, March 2006.

[RFC4378] Allan, D. and T. Nadeau, "A Framework for Multi-Protocol Label Switching (MPLS) Operations and Management (OAM)", RFC 4378, March 2006.

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

[RFC4447] Martini, L., Rosen, E., El-Aawar, N., Smith, T., and G. Heron, "Pseudowire Setup and Maintenance Using the Label Distribution Protocol (LDP)", RFC 4447, April 2006.

[RFC4805] Nicklass, O., Ed., "Definitions of Managed Objects for the DS1, J1, E1, DS2, and E2 Interface Types", RFC 4805, March 2007.

[RFC5085] Nadeau, T. and C. Pignataro, "Pseudowire Virtual Circuit Connectivity Verification (VCCV): A Control Channel for Pseudowires", RFC 5085, December 2007.

[RFC5601] Nadeau, T., Ed. and D. Zelig, Ed. "Pseudowire (PW) Management Information Base (MIB)", RFC 5601, July 2009.

[RFC5602] Zelig, D., Ed., and T. Nadeau, Ed., "Pseudowire (PW) over MPLS PSN Management Information Base (MIB)", RFC 5602, July 2009.

[RFC5654] Niven-Jenkins, B., et al, "MPLS-TP Requirements", RFC5654, September 2009.

[RFC5880] Katz, D. and D. Ward, "Bidirectional Forwarding Detection", 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)", RFC5885, June 2010.

[RFC5950] Gray, E., Mansfield, S., Lam, K., "MPLS-TP Network Management Framework", RFC 5950, September 2010.

[RFC5951] Gray, E., Mansfield, S., Lam, K., "MPLS TP Network Management Requirements", RFC 5951, September 2010.

[RFC6370] Bocci, M., Swallow, G., and E. Gray, "MPLS Transport Profile (MPLS-TP) Identifiers", RFC 6370, September 2011.

[MPLS-TP-OAM-FWK] Busi, I. and B. Niven-Jenkins, "MPLS-TP OAM Framework and Overview", 2009, <draft-ietf-mpls-tp-oam-framework>.

12. Authors' Addresses

Daniel King
Old Dog Consulting
UK
Email: daniel@olddog.co.uk

Venkatesan Mahalingam
Aricent
India
Email: venkat.mahalingams@gmail.com

Adrian Farrel
Old Dog Consulting
UK
Email: adrian@olddog.co.uk

Scott Mansfield
Ericsson
300 Holger Way, San Jose, CA 95134, US
Phone: +1 724 931 9316
Email: scott.mansfield@ericsson.com

Jeong-dong Ryoo
ETRI
161 Gajeong, Yuseong, Daejeon, 305-700, South Korea
Phone: +82 42 860 5384
Email: ryoo@etri.re.kr

A S Kiran Koushik
Cisco Systems Inc.
Email: kkoushik@cisco.com

A. Karmakar
Cisco Systems Inc.
Email: akarmaka@cisco.com

Sam Aldrin
Huawei Technologies, co.
2330 Central Express Way,
Santa Clara, CA 95051, USA
Email: aldrin.ietf@gmail.com

Network Working Group
Internet-Draft
Intended status: Informational
Expires: October 18, 2012

N. Sprecher
Nokia Siemens Networks
L. Fang
Cisco

April 17, 2012

An Overview of the OAM Tool Set for MPLS based Transport Networks
draft-ietf-mpls-tp-oam-analysis-09.txt

Abstract

This document provides an overview of the OAM toolset for MPLS based Transport Networks (MPLS-TP). The toolset consists of a comprehensive set of fault management and performance monitoring capabilities (operating in the data-plane) which are appropriate for transport networks as required in RFC 5860 and support the network and services at different nested levels. This overview includes a brief recap of MPLS-TP OAM requirements and functions, and of generic mechanisms created in the MPLS data plane to allow the OAM packets run in-band and share their fate with data packets. The protocol definitions for each of the MPLS-TP OAM tools are defined in separate documents (RFCs or Working Group drafts) which are referenced by this document.

Status of this Memo

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

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

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

This Internet-Draft will expire on October 18, 2012.

Copyright Notice

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

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

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

Table of Contents

1. Introduction

1.1. Scope

The MPLS Transport Profile (MPLS-TP) architectural framework is defined in [RFC 5921], and it describes common set of protocol functions that supports the operational models and capabilities typical of such networks.

OAM (Operations, Administration, and Maintenance) plays a significant role in carrier networks, providing methods for fault management and performance monitoring in both the transport and the service layers in order to improve their ability to support services with guaranteed and strict Service Level Agreements (SLAs) while reducing their operational costs.

[RFC 5654], in general, and [RFC 5860], in particular, define a set of requirements for OAM functionality for MPLS-Transport Profile (MPLS-TP)Label Switched Paths (LSPs)), Pseudowires (PWs) and sections.

The OAM solution, developed by the joint IETF and ITU-T MPLS-TP project, has three objectives:

- The OAM toolset should be developed based on existing MPLS architecture, technology, and toolsets.
- The OAM operational experience should be similar to that in other transport networks.
- The OAM toolset developed for MPLS based transport networks needs to be fully inter-operable with existing MPLS OAM tools as documented in [RFC 5860].

The MPLS-TP OAM toolset is based on the following existing tools:

- LSP-Ping as defined in [RFC 4379].
- Bidirectional Forwarding Detection (BFD) as defined in [RFC 5880] and refined in [RFC 5884].
- ITU-T OAM for Ethernet toolset as defined in [Y.1731]. This has been used for functionality guidelines for the performance measurement tools that were not previously supported in MPLS.

It should be noted that certain extensions and adjustments have been specified relative to the existing MPLS tools, in order to conform to the transport environment and the requirements of MPLS-TP. However,

compatibility with the existing tools has been maintained.

This document provides an overview of the MPLS-TP OAM toolset, which consists of tools for MPLS-TP fault management and performance monitoring. This overview includes a brief recap of MPLS-TP OAM requirements and functions, and of the generic mechanisms used to support the MPLS-TP OAM operation.

The protocol definitions for each individual MPLS-TP OAM tool are specified in separate RFCs (or Working Group documents while this document is work in progress), which are referenced by this document.

In addition, the document includes a table that cross-references the solution documents to the OAM functionality supported. Finally, the document presents the applicability and utilization of each tool in the MPLS-TP OAM toolset.

1.2. Contributing Authors

Elisa Bellagamba	Ericsson
Yaacov Weingarten	Nokia Siemens Networks
Dan Frost	Cisco
Nabil Bitar	Verizon
Raymond Zhang	Alcatel Lucent
Lei Wang	Telenor
Kam Lee Yap	XO Communications
John Drake	Juniper
Yaakov Stein	RAD
Anamaria Fulignoli	Ericsson
Italo Busi	Alcatel Lucent
Huub van Helvoort	Huawei
Thomas Nadeau	Computer Associate
Henry Yu	TW Telecom
Mach Chen	Huawei
Manuel Paul	Deutsche Telekom

1.3. Acronyms

This document uses the following acronyms:

ACH	Associated Channel Header
AIS	Alarm Indication Signal
BFD	Bidirectional Forwarding Detection
CC-CV	Continuity Check and Connectivity Verification
DM	Delay Measurement
FM	Fault Management
G-ACh	Generic Associated Channel
GAL	G-ACh Label
GMPLS	Generalized Multi-Protocol Label Switching
IANA	Internet Assigned Names Authority
LDI	Link Down Indication
LKR	Lock Report
LM	Loss Measurement
LOC	Loss of Continuity
LSP	Label Switched Path
MEP	Maintenance Entity Group End Point
MEG	Maintenance Entity Group
MIP	Maintenance Entity Group Intermediate Point
MPLS	MultiProtocol Label Switching
MPLS-TP	Transport Profile for MPLS
OAM	Operations, Administration, and Maintenance
PM	Performance Monitoring
PW	Pseudowire
RDI	Remote Defect Indication
SLA	Service Level Agreement
TLV	Type, Length, Value
VCCV	Virtual Circuit Connectivity Verification

2. Basic OAM Infrastructure Functionality

[RFC 5860] defines a set of requirements on OAM architecture and general principles of operations, which are evaluated below:

[RFC 5860] requires that --

- OAM mechanisms in MPLS-TP are independent of the transmission media and of the client service being emulated by the PW ([RFC 5860], section 2.1.2).

- MPLS-TP OAM must be able to support both an IP based and non-IP based environment. If the network is IP based, i.e. IP routing and forwarding are available, then it must be possible to choose to make use of IP capabilities. On the other hand, in

environments where IP functionality is not available, the OAM tools must still be able to operate independent of IP forwarding and routing ([RFC 5860], section 2.1.4). It is required to have OAM interoperability between distinct domains materializing the environments ([RFC 5860], section 2.1.5).

- o all OAM protocols support identification information, at least in the form of IP addressing structure and be extensible to support additional identification schemes ([RFC 5860], section 2.1.4).

- o OAM packets and the user traffic are congruent (i.e. OAM packets are transmitted in-band) and there is a need to differentiate OAM packets from user-plane packets ([RFC 5860], section 2.1.3). Inherent in this requirement is the principle that full operation of the MPLS-TP OAM must be possible independently of the control or management plane used to operate the network ([RFC 5860], section 2.1.3).

- o MPLS-TP OAM supports point-to-point bidirectional PWs, point-to-point co-routed bidirectional LSPs, point-to-point bidirectional Sections ([RFC 5860], section 2.1.1). The applicability of particular MPLS-TP OAM functions to point-to-point associated bidirectional LSPs, point-to-point unidirectional LSPs, and point-to-multipoint LSPs, is described in ([RFC 5860], section 2.2)). In addition, MPLS-TP OAM supports these LSPs and PWs when they span either a single or multiple domains ([RFC 5860], section 2.1.1).

- o OAM packets may be directed to an intermediate point of a LSP/PW ([RFC 5860], sections 2.2.3, 2.2.4 and 2.2.5).

[RFC 5860] recommends that any protocol solution, meeting one or more functional requirement(s), be the same for PWs, LSPs, and Sections (section 2.2).

The following document-set addresses the basic requirements listed above:

- o The [RFC 6371] document describes the architectural framework for conformance to the basic requirements listed above. It also defines the basic relationships between the MPLS structures, e.g. LSP, PW, and the structures necessary for OAM functionality, i.e. the Managed Entity Group, its End-points, and Intermediate Points.

- o The [RFC 5586] document specifies the use of the MPLS-TP in-band control channels. It generalizes the applicability of the Pseudowire (PW) Associated Channel Header (ACH) to MPLS LSPs and Sections, by defining a Generic Associated Channel (G-ACh). The

G-ACh allows control packets to be multiplexed transparently over LSPs and sections, similar to that of PW VCCV [RFC 5085]. The Generic Association Label (GAL) is defined by assigning a reserved MPLS label value and is used to identify the OAM control packets. The value of the ACH Channel Type field indicates the specific protocol carried on the associated control channel. Each MPLS-TP OAM protocol has an IANA assigned channel type allocated to it.

[RFC 5085] defines an Associated Channel Header (ACH) which provides a PW associated control channel between a PW's endpoints, over which OAM and other control messages can be exchanged. [RFC 5586] generalizes MPLS-TP generalized the PW Associated Channel Header (ACH) to provide common in-band control channels also at the LSP and MPLS-TP link levels. The G-ACh allows control packets to be multiplexed transparently over the same LSP or MPLS-TP link as in PW VCCV. Multiple control channels can exist between end points.

[RFC 5085] also defines a label-based exception mechanism that helps an LSR to identify the control packets and direct them to the appropriate entity for processing. The use of G-ACh and GAL provides the necessary mechanisms to allow OAM packets run in-band and share their fate with data packets. It is expected that all of the OAM protocols will be used in conjunction with this Generic Associated Channel.

- The [RFC 6370] document provides an IP-based identifier set for MPLS-TP that can be used to identify the transport entities in the network and referenced by the different OAM protocols.
  [MPLS TP ITU Idents] augments that set of identifiers to include identifier information in a format used by the ITU-T. Other identifier sets may be defined as well.

3. MPLS-TP OAM Functions

The following sections discuss the OAM functions that are required in [RFC 5860] and expanded upon in [RFC 6371].

3.1. Continuity Check and Connectivity Verification

Continuity Check and Connectivity Verification (CC-CV) are OAM operations generally used in tandem, and complement each other. These functions are generally run proactively, but may also be used on-demand for diagnoses of a specific condition. Proactively [RFC 5860] states that the function should allow the MEPs to monitor the liveliness and connectivity of a transport path (LSP, PW or a section) between them. In on-demand mode, this function should

support monitoring between the MEPs and, in addition, between a MEP and MIP. Note that as specified in sections 3.3 and 3.4 of [RFC 6371], a MEP and a MIP can reside in an unspecified location within a node, or in a particular interface on a specific side of the forwarding engine.

The [RFC 6371] highlights the need for the CC-CV messages to include unique identification of the MEG that is being monitored and the MEP that originated the message. The function, both proactively and in on-demand mode, needs to be transmitted at regular transmission rates pre-configured by the operator.

3.1.1. Documents for CC-CV tools

[RFC 6428] defines BFD extensions to support proactive CC-CV applications.

[RFC 6426] provides LSP-Ping extensions that are used to implement on-demand Connectivity Verification.

Both of these tools will be used within the framework of the basic tools described above, in section 2.

3.2. Remote Defect Indication

Remote Defect Indication (RDI) is used by a path end-point to report that a defect is detected on a bi-directional connection to its peer end-point. [RFC 5860] points out that this function may be applied to a unidirectional LSP only if a return path exists. [RFC 6371] points out that this function is associated with the proactive CC-CV function.

3.2.1. Documents for RDI

The [RFC 6428] document includes an extension for BFD that would include the RDI indication in the BFD format, and a specification of how this indication is to be used.

3.3. Route Tracing

[RFC 5860] defines that there is a need for functionality that would allow a path end-point to identify the intermediate (if any) and end-points of the path (LSP, PW or a section). This function would be used in on-demand mode. Normally, this path will be used for bidirectional PW, LSP, and sections, however, unidirectional paths may be supported only if a return path exists.

3.3.1. Documents for Route Tracing

The [RFC 6426] document that specifies the LSP-Ping enhancements for MPLS-TP on-demand Connectivity Verification includes information on the use of LSP-Ping for route tracing of a MPLS-TP transport path.

3.4. Alarm Reporting

Alarm Reporting is a function used by an intermediate point of a path (LSP or PW), that becomes aware of a fault on the path, to report to the end-points of the path. [RFC 6371] states that this may occur as a result of a defect condition discovered at a server layer. The intermediate point generates an Alarm Indication Signal (AIS) that continues until the fault is cleared. The consequent action of this function is detailed in [RFC 6371].

3.4.1. Documents for Alarm Reporting

MPLS-TP defines a new protocol to address this functionality that is documented in [RFC 6427]. This protocol uses all of the basic mechanisms detailed in Section 2.

3.5. Lock Instruct

The Lock Instruct function is an administrative control tool that allows a path end-point to instruct its peer end-point to lock the path (LSP, PW or section). The tool is necessary to support single-side provisioning for administrative locking, according to [RFC 6371]. This function is used on-demand.

3.5.1. Documents for Lock Instruct

The [RFC 6435] document describes the details of a new ACH based protocol format for this functionality.

3.6. Lock Reporting

Lock reporting, defined in [RFC 5860], is similar to the Alarm Reporting function described above. It is used by an intermediate point to notify the end points of a transport path (LSP or PW) that an administrative lock condition exists for this transport path.

3.6.1. Documents for Lock Reporting

MPLS-TP defines a new protocol to address this functionality that is documented in [RFC 6427]. This protocol uses all of the basic mechanisms detailed in Section 2.

3.7. Diagnostic

The [RFC 5860] indicates that there is need to provide a OAM function that would enable conducting different diagnostic tests on a PW, LSP, or Section. The [RFC 6371] provides two types of specific tests to be used through this functionality:

- Throughput Estimation - allowing the provider to verify the bandwidth/throughput of a transport path. This is an out-of-service tool, that uses special packets of varying sizes to test the actual bandwidth and/or throughput of the path.

- Data-plane loopback - this out-of-service tool causes all traffic that reaches the target node, either a MEP or MIP, to be looped back to the originating MEP. For targeting MIPs, a co-routed bi-directional path is required.

3.7.1. Documents for Diagnostic Testing

The [RFC 6435] document describes the details of a new ACH based protocol format for the Data-plane loopback functionality.

The tool for Throughput Estimation tool is under study.

3.8. Packet Loss Measurement

Packet Loss Measurement is required by [RFC 5860] to provide a quantification of the packet loss ratio on a transport path. This is the ratio of the number of user packets lost to the total number of user packets during a defined time interval. To employ this function, [RFC 6371] defines that the two end-points of the transport path should exchange counters of messages transmitted and received within a time period bounded by loss-measurement messages. The framework warns that there may be small errors in the computation that result from various issues.

3.8.1. Documents for Packet Loss Measurement

The [RFC 6374] document describes the protocol formats and procedures for using the tool and enable efficient and accurate measurement of packet loss, delay, and throughput in MPLS networks. [RFC 6375] describes a profile of the general MPLS loss, delay, and throughput measurement techniques that suffices to meet the specific requirements of MPLS-TP. Note that the tool logic is based on the behavior of the parallel function described in [Y.1731].

3.9. Packet Delay Measurement

Packet Delay Measurement is a function that is used to measure one-way or two-way delay of a packet transmission between a pair of the end-points of a path (PW, LSP, or Section), as described in [RFC 5860]. Where:

- One-way packet delay is the time elapsed from the start of transmission of the first bit of the packet by a source node until the reception of the last bit of that packet by the destination node.

- Two-way packet delay is the time elapsed from the start of transmission of the first bit of the packet by a source node until the reception of the last bit of the loop-backed packet by the same source node, when the loopback is performed at the packet's destination node.

[RFC 6371] describes how the tool could be performed (both in proactive and on-demand modes) for either one-way or two-way measurement. However, it warns that the one-way delay option requires precise time synchronization between the end-points.

3.9.1. Documents for Delay Measurement

The [RFC 6374] document describes the protocol formats and procedures for using the tool and enable efficient and accurate measurement of packet loss, delay, and throughput in MPLS networks. [RFC 6375] describes a profile of the general MPLS loss, delay, and throughput measurement techniques that suffices to meet the specific requirements of MPLS-TP. Note that the tool logic is based on the behavior of the parallel function described in [Y.1731].

4. MPLS-TP OAM documents guide

The complete MPLS-TP OAM protocol suite is covered by a small set of existing IETF documents. This set of documents may be expanded in the future to cover additional OAM functionality. In order to allow the reader to understand this set of documents, a cross-reference of the existing documents (IETF RFCs or Internet drafts while this document is work in progress) for the initial phase of the specification of MPLS based transport networks is provided below.

[RFC 5586] provides a specification of the basic structure of protocol messages for in-band data plane OAM in an MPLS environment.

[RFC 6370] provides definitions of different formats that may be used

within OAM protocol messages to identify the network elements of a MPLS based transport network.

The following table (Table 1) provides the summary of proactive MPLS-TP OAM Fault Management toolset functions, associated tool/ protocol, and the corresponding IETF RFCs where they are defined.

OAM Functions	OAM Tools/Protocols	RFCs
Continuity Check and Connectivity Verification	Bidirectional Forwarding Detection (BFD)	[RFC 6428]
Remote Defect Indication (RDI)	Flag in Bidirectional Forwarding Detection (BFD) message	[RFC 6428]
Alarm Indication Signal (AIS)	G-ACh bases AIS message	[RFC 6427]
Link Down Indication (LDI)	Flag in AIS message	[RFC 6427]
Lock Reporting (LKR)	G-ACh bases LKR message	[RFC 6427]

Proactive Fault Management OAM Toolset

Table 1

The following table (Table 2) provides an overview of the on-demand MPLS-TP OAM Fault Management toolset functions, associated tool/ protocol, and the corresponding IETF RFCs they are defined.

OAM Functions	OAM Tools/Protocols	RFCs
Connectivity Verification	LSP Ping	[RFC 6426]
Diagnostic: Loopback and Lock Instruct	(1) G-ACh based Loopback and Lock Instruct, (2) LSP Ping	[RFC 6435]

Lock Instruct(LI)	Flag in AIS message	[RFC 6427]

On Demand Fault Management OAM Toolset

Table 2

The following table (Table 3) provides the Performance Monitoring Fuctions, asscociated tool/protocol definitions, and corresponding RFCs.

OAM Functions	OAM Tools/Protocols	RFCs
Packet Loss Measurement (LM)	G-ACh based LM & DM query messages	[RFC 6374] [RFC 6375]
Packet Delay Measurement (DM)	G-ACh based LM & DM query messages	[RFC 6374] [RFC 6375]
Throughput Measurement	derived from Loss Measurement	[RFC 6374] [RFC 6375]
Delay Variation Measurement	derived from Delay Measurement	[RFC 6374] [RFC 6375]

Performance Monitoring OAM Toolset

Table 3

5. OAM Toolset Applicability and Utilization

The following subsections present the MPLS-TP OAM toolset from the perspective of the specified protocols and identifies which of the required functionality is supported by the particular protocol.

5.1. Connectivity Check and Connectivity Verification

Proactive Continuity Check and Connectivity Verification (CC-CV) functions are used to detect loss of continuity (LOC), and unintended connectivity between two MEPs. Loss of connectivity, mis-merging, mis-connectivity, or unexpected Maintenance Entity Group End Points (MEPs) can be detected using the CC-CV tools. See Section 3.1, 3.2, 3.3 in this document for CC-CV protocol references.

The CC-CV tools are used to support MPLS-TP fault management, performance management, and protection switching. Proactive CC-CV control packets are sent by the source MEP to sink MEP. The sink MEP monitors the arrival of the CC-CV control packets and detects the defect. For bidirectional transport paths, the CC-CV protocol is, usually, transmitted simultaneously in both directions.

The transmission interval of CC-CV control packet can be configured. For example:

- 3.3ms is the default interval for protection switching.
- 100ms is the default interval for performance monitoring.
- 1s is the default interval for fault management.

5.2. Diagnostic Tests and Lock Instruct

[RFC 6435] describes a protocol that provides a mechanism is provided to Lock and unlock traffic (e.g. data and control traffic) or specific OAM traffic at a specific LSR on the path of the MPLS-TP LSP to allow loop back of the traffic to the source.

These diagnostic functions apply to associated bidirectional MPLS-TP LSPs, including MPLS-TP LSPs, bi-directional RSVP-TE tunnels (which is relevant for MPLS-TP dynamic control plane option with GMPLS), and single segment and multi-segment pseudowires. [RFC 6435] provides the protocol definition for diagnostic tests functions.

The Lock operation instruction is carried in an MPLS Loopback request message sent from a MEP to a trail-end MEP of the LSP to request that the LSP be taken out of service. In response, the Lock operation reply is carried in a Loopback response message sent from the trail-end MEP back to the originating MEP to report the result.

The loopback operations include:

- Lock: take an LSP out of service for maintenance.
- Unlock: Restore a previously locked LSP to service.
- Set_Full_Loopback and Set_OAM_Loopback
- Unset_Full_Loopback and Set_OAM_Loopback

Operators can use the loopback mode to test the connectivity or performance (loss, delay, delay variation, and throughput) of given LSP up to a specific node on the path of the LSP.

5.3. Lock Reporting

The Lock Report (LKR) function is used to communicate to the client (sub-) layer MEPs the administrative locking of a server (sub-) layer MEP, and consequential interruption of data traffic forwarding in the client (sub-) layer. See Section 3.6 in this document for Lock Reporting protocol references.

When operator is taking the LSP out of service for maintenance or other operational reason, using the LKR function can help to distinguish the condition as administrative locking from defect condition.

The Lock Report function would also serve the purpose of alarm suppression in the MPLS-TP network above the level at which the Lock has occurred. The receipt of an LKR message may be treated as the equivalent of loss of continuity at the client layer.

5.4. Alarm Reporting and Link Down Indication

Alarm Indication Signal (AIS) message serves the purpose of alarm suppression upon the failure detection in the server (-sub) layer. When the Link Down Indication (RDI) is set, the AIS message may be used to trigger recovery mechanisms.

When a server MEP detects the failure, it asserts Loss of Continuity (LOC) or signal fail which sets the flag up to generate OAM packet with AIS message. The AIS message is forwarded to downstream sink MEP in the client layer. This would enable the client layer to suppress the generation of secondary alarms.

A Link Down Indication (LDI) flag is defined in the AIS message. The LDI flag is set in the AIS message in response to detecting a fatal failure in the server layer. Receipt of an AIS message with this flag set may be interpreted by a MEP as an indication of signal fail at the client layer.

The protocols for Alarm Indication Signal (AIS) and Link Down Indication (LDI) are defined in [RFC 6427].

Fault OAM messages are generated by intermediate nodes where an LSP is switched, and propagated to the end points (MEPs).

From a practical point of view, when both proactive Continuity Check functions and LDI are used, one may consider running the proactive Continuity Check functions at a slower rate (e.g. longer BFD hello intervals), and reply on LDI to trigger fast protection switch over upon failure detection in a given LSP.

5.5. Remote Defect Indication

Remote Defect Indication (RDI) function enables an End Point to report to the other End Point that a fault or defect condition is detected on the PW, LSP, or Section for which they are the End Points.

The RDI OAM function is supported by the use of Bidirectional Forwarding Detection (BFD) Control Packets [RFC 6428]. RDI is only used for bidirectional connections and is associated with proactive CC-CV activation.

When an end point (MEP) detects a signal failure condition, it sets the flag up by setting the diagnostic field of the BFD control packet to a particular value to indicate the failure condition on the associated PW, LSP, or Section, and transmitting the BFD control packet with the failure flag up to the other end point (its peer MEP).

The RDI function can be used to facilitate protection switching by synchronizing the two end points when unidirectional failure occurs and is detected by one end.

5.6. Packet Loss and Delay Measurement

The packet loss and delay measurement toolset enables operators to measure the quality of the packet transmission over a PW, LSP, or Section. Section 3.8 in this document defined the protocols for packet loss measurement and 3.9 in defined the protocols for packet delay measurement.

The loss and delay protocols have the following characteristics and capabilities:

- They support measurement of packet loss, delay and throughput over Label Switched Paths (LSPs), pseudowires, and MPLS sections.

- The same LM and DM protocols can be used for both continuous/ proactive and selective/on-demand measurement.

- The LM and DM protocols use a simple query/response model for bidirectional measurement that allows a single node - the querier - to measure the loss or delay in both directions.

- The LM and DM protocols use query messages for unidirectional loss and delay measurement. The measurement can either be carried out at the downstream node(s) or at the querier if an out-of-band return path is available.

- The LM and DM protocols do not require that the transmit and receive interfaces be the same when performing bidirectional measurement.
- The LM supports test-message-based measurement (i.e. inferred mode) as well as measurement based on data-plane counters (i.e. direct mode).
- The LM protocol supports both 32-bit and 64-bit counters.
- The LM protocol supports measurement in terms of both packet counts and octet counts although for simplicity only packet counters are currently included in the MPLS-TP profile.
- The LM protocol can be used to measure channel throughput as well as packet loss.
- The DM protocol supports varying the measurement message size in order to measure delays associated with different packet sizes.
- The DM protocol uses IEEE 1588 timestamps by default but also supports other timestamp formats such as NTP.

6. IANA Considerations

This document makes no request of IANA.

The OAM tools and functions defined under G-ACh use IANA assigned code points. the codes are defined in the corresponding IETF RFCs

Note to RFC Editor:

this section may be removed on publication as an RFC.

7. Security Considerations

This document as an overview of MPLS OAM tools does not by itself raise any particular security considerations.

The general security considerations are provided in [RFC 6920] and [MPLS-TP Security Frwk]. Security considerations for each function in the OAM toolset have been documented in each document that specifies the particular functionality.

OAM in general is always an area where the security risk is high, e.g. confidential information may be intercepted for attackers to

again access to the networks, therefore authentication, authorization, and encryption need to be enforced for prevent security breach.

In addition to implement security protocol, tools, and mechanisms, following strict operation security procedures is very important, especially MPSL-TP static provisioning processes involve operator direct interactions with NMS and devices, its critical to prevent human errors and malicious attacks.

Since MPLS-TP OAM uses G-ACh, the security risks and mitigation described in [RFC 5085] apply here. In short, the G-ACh could be intercepted, or false G-ACh packets could be inserted. DoS attack could happen by flooding G-ACh messages to peer devices. To mitigate this type of attacks, throttling mechanisms can be used. For more details, please see [RFC 5085].

8. Acknowledgements

The authors would like to thank the MPLS-TP experts from both the IETF and ITU-T for their helpful comments. In particular, we would like to thank Loa Andersson, and the Area Directors for their suggestions and enhancements to the text.

Thanks to Tom Petch for useful comments and discussions.

Thanks to Rui Costa for his review and comments which helped improve this doecument.

9. References

9.1. Normative References

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

[RFC 5085]
Nadeau, T. and C. Pignataro, "Pseudowire Virtual Circuit Connectivity Verification (VCCV): A Control Channel for Pseudowires", RFC 5085, December 2007.

[RFC 5586]
Bocci, M., Bryant, S., and M. Vigoureux, "MPLS Generic Associated Channel", RFC 5586, June 2009.

[RFC 5654]
Niven-Jenkins, B., Nadeau, T., and C. Pignataro, "Requirements for the Transport Profile of MPLS", RFC 5654, April 2009.

[RFC 5860]
Vigoureux, M., Betts, M., and D. Ward, "Requirements for OAM in MPLS Transport Networks", RFC 5860, April 2009.

[RFC 5880]
Katz, D. and D. Ward, "Bidirectional Forwarding Detection", RFC 5880, February 2009.

[RFC 5884]
Aggarwal, R., Kompella, K., Nadeau, T., and G. Swallow, "BFD For MPLS LSPs", RFC 5884, June 2008.

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

[RFC 6370]
Bocci, M., Swallow, G., and E. Gray, "MPLS-TP Identifiers", RFC 6370, September 2011.

[RFC 6371]
Busi, I., Niven-Jenkins, B., and D. Allan, "MPLS-TP OAM Framework and Overview", RFC 6371, September 2011.

[RFC 6374]
Frost, D. and S. Bryant, "Packet Loss and Delay Measurement for MPLS Networks", RFC 6374, September 2011.

[RFC 6375]
Frost, D. and S. Bryant, "A Packet Loss and Delay Measurement Profile for MPLS-based Transport Networks", RFC 6375, September 2011.

[RFC 6426]
Bahadur, N., Aggarwal, R., Boutros, S., and E. Gray, "MPLS on-demand Connectivity Verification, Route Tracing and Adjacency Verification", RFC 6426, August 2011.

[RFC 6427]
Swallow, G., Fulignoli, A., and M. Vigoureux, "MPLS Fault Management OAM", RFC 6427, September 2011.

[RFC 6428]
Allan, D. and G. Swallow, "Proactive Connectivity Verification, Continuity Check and Remote Defect indication for MPLS Transport Profile", RFC 6428, August 2011.

[RFC 6435]
Boutros, S., Sivabalan, S., Aggarwal, R., Vigoureux, M., and X. Dai, "MPLS Transport Profile Lock Instruct and Loopback Functions", RFC 6435, September 2011.

9.2. Informative References

[MPLS-TP Security Frwk]
Fang, L., Niven-Jenkins, B., and S. Mansfield, "MPLS-TP Security Framework", ID draft-ietf-mpls-tp-security-framework-02, May 2011.

[RFC 6920]
Fang, L., "Security Framework for MPLS and GMPLS Networks", RFC 5920, July 2010.

[Y.1731] International Telecommunications Union - Standardization, "OAM functions and mechanisms for Ethernet based networks", ITU Y.1731, May 2006.

[MPLS TP ITU Idents]
Winter, R., van Helvoort, H., and M. Betts, "MPLS-TP Identifiers Following ITU-T Conventions", ID draft-ietf-mpls-tp-itu-t-identifiers-02, July 2011.

Authors' Addresses

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

Email: nurit.sprecher@nsn.com

Luyuan Fang
Cisco
111 Wood Avenue South
Iselin, NJ 08830
USA

Email: lufang@cisco.com

MPLS Working Group
Internet Draft
Intended status: Informational
Expires: April 2014

H. van Helvoort (Ed)
Huawei Technologies

L. Andersson (Ed)
Huawei Technologies

N. Sprecher (Ed)
Nokia Solutions and Networks

October 20, 2013

A Thesaurus for the Terminology used in Multiprotocol Label Switching Transport Profile (MPLS-TP) drafts/RFCs and ITU-T's Transport Network Recommendations.
draft-ietf-mpls-tp-rosetta-stone-13

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/1id-abstracts.txt.

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

This Internet-Draft will expire on April 20, 2014.

Copyright Notice

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

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

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

Abstract

MPLS Transport Profile (MPLS-TP) is based on a profile of the MPLS and Pseudowire (PW) procedures as specified in the MPLS-TE, PW and Multi-Segment Pseudowire (MS-PW) architectures developed by the Internet Engineering Task Force (IETF). The International Telecommunications Union Telecommunications Standardization Sector (ITU-T) has specified a Transport Network architecture.

This document provides a thesaurus for the interpretation of MPLS-TP terminology within the context of the ITU-T Transport Network Recommendations.

It is important to note that MPLS-TP is applicable in a wider set of contexts than just Transport Networks. The definitions presented in this document do not provide exclusive nor complete interpretations of MPLS-TP concepts. This document simply allows the MPLS-TP terms to be applied within the Transport Network context.

Table of Contents

1 Introduction

Multiprotocol Label Switching - Transport Profile (MPLS-TP) has been developed by the IETF to facilitate the Operation, Administration and Management of Label Switched Paths (LSPs) to be used in a Transport Network environment as defined by the ITU-T.

The ITU-T has specified a Transport Network architecture for the transfer of signals from different technologies. This architecture forms the basis of many Recommendations within the ITU-T.

Because of the difference in historic background of MPLS, and inherently MPLS-TP (the Internet) and the Transport Network (ITU Telecommunication Sector), the terminology used is different.

This document provides a thesaurus for the interpretation of MPLS-TP terminology within the context of the ITU-T Transport Network Recommendations. This allows MPLS-TP documents to be generally understood by those familiar with MPLS RFCs. The definitions presented in this document do not provide exclusive or complete interpretations of the ITU-T Transport Network concepts.

1.1 Contributing Authors

Italo Busi, Ben Niven-Jenkins, Enrique Hernandez-Valencia, Lieven Levrau, Dinesh Mohan, Stuart Bryant, Dan Frost, Matthew Bocci, Vincenzo Sestito, Vigoureux, Yaacov Weingarten

1.2 Abbreviations

CE Customer Edge

DCC Data Communication Channel

DCN Data Communication Network

ECC Embedded Communication Channel

EMF Equipment Management Function

EMS Element Management System

GAL Generic Associated Channel Label

NEF Network Element Function

LER Label Edge Router

LSR Label Switching Router

MCC Management Communication Channel

MCN Management Communication Network

ME Maintenance Entity

MEG Maintenance Entity Group

MEP Maintenance Entity Group End Point

MIP Maintenance Entity Group Intermediate Point

MPLS Multiprotocol Label Switching

MPLS-TP MPLS Transport Profile

MS-PW Multi-Segment Pseudowire

NE Network Element

OAM Operations, Administration, and Maintenance

OSS Operations Support System

PM Performance Monitoring

PST Path Segment Tunnel

PW Pseudowire

S-PE PW Switching Provider Edge

SCC Signaling Communication Channel

SCN Signaling Communication Network

SPME Sub-Path Maintenance Element

T-PE PW Terminating Provider Edge

TCM Tandem Connection Monitoring

2 Terminology

2.1 MPLS-TP Terminology Sources

MPLS-TP terminology is principally defined in [RFC3031]. Other documents provide further key definitions including [RFC4397].

2.2 ITU-T Transport Network Terminology Sources

The ITU-T Transport Network is specified in a number of Recommendations: generic functional architectures and requirements are specified in [ITU-T_G.805], [ITU-T_G.806], and [ITU-T_G.872]. ITU-T Recommendation [ITU-T_G.8101] contains an overview of the Terms and Definitions for transport MPLS.

2.3 Common Terminology Sources

The work in this document builds on the shared view of MPLS requirements. It is intended to provide a source for common MPLS-TP terminology. In general the original terminology is used.

The following sources are used:
IETF framework and requirements RFCs: [RFC6371], [RFC6372], [RFC5654], [RFC5921], [RFC5860], [RFC5951], [RFC3031] and [RFC4397].
ITU-T architecture and requirements Recommendations: [ITU-T_G.8101], [ITU-T_G.805], [ITU-T_G.806], [ITU-T_G.872], [ITU-T G.7710] and [ITU-T Y.2611].

3 Thesaurus

3.1 Associated bidirectional path:

A path that supports traffic flow in both directions but that is constructed from a pair of unidirectional paths (one for each direction) that are associated with one another at the path's ingress/egress points. An associated bidirectional path needs not be a single management and operational entity. The forward and backward directions are setup, monitored, and protected

independently. As a consequence, they may or may not follow the same route (links and nodes) across the network.

3.2 Bidirectional path:

A path that supports traffic flow in two opposite directions, i.e. the forward and backward direction.

3.3 Client layer network:

In a client/server relationship (see [ITU-T_G.805]), the client layer network receives a (transport) service from the lower server layer network (usually the layer network under consideration).

3.4 Communication Channel:

A logical channel between network elements (NEs) that can be used - e.g. - for management plane application or control plane applications. The physical channel supporting the Communication Channel is technology specific. See [RFC5951] Appendix A.

3.5 Concatenated Segment:

A serial-compound link connection as defined in [ITU-T_G.805]. A concatenated segment is a contiguous part of an LSP or MS-PWthat comprises a set of segments and their interconnecting nodes in sequence. See also "Segment".

3.6 Control Plane:

Within the scope of [RFC5654], the control plane performs transport path control functions. Through signalling, the control plane sets up, modifies and releases transport paths, and may recover a transport path in case of a failure. The control plane also performs other functions in support of transport path control, such as routing information dissemination. It is possible to operate an MPLS-TP network without using a Control Plane.

3.7 Co-routed bidirectional path:

A path where the forward and backward directions follow the same route (links and nodes) across the network. A co-routed bidirectional path is managed and operated as a single entity. Both directions are setup, monitored and protected as a single entity. A transport network path is typically co-routed.

3.8 Data Communication Network (DCN):

A network that supports Layer 1 (physical layer), Layer 2 (data-link layer), and Layer 3 (network layer) functionality for distributed management communications related to the management plane, for distributed routing and signaling communications related to the control plane, and other operations communications (e.g., order-wire/voice communications, software downloads, etc.).

3.9 Defect:

The situation for which the density of anomalies has reached a level where the ability to perform a required function has been interrupted. Defects are used as input for Performance Monitoring (PM), the control of consequent actions, and the determination of fault cause. See also [ITU-T_G.806].

3.10 Domain:

A domain represents a collection of entities (for example network elements) that are grouped for a particular purpose, examples of which are administrative and/or managerial responsibilities, trust relationships, addressing schemes, infrastructure capabilities, aggregation, survivability techniques, distributions of control functionality, etc. Examples of such domains include IGP areas and Autonomous Systems.

3.11 Embedded Communication Channel (ECC):

A logical operations channel between network elements (NEs) that can be utilized by multiple applications (e.g., management plane applications, control plane applications, etc.). The physical channel supporting the ECC is technology specific. An example of a physical channel supporting the ECC is a Data Communication Channel (DCC) within SDH.

3.12 Equipment Management Function (EMF):

The equipment management function (EMF) provides the means through which an element management system (EMS) and other managing entities manage the network element function (NEF). See [ITU-T G.7710].

3.13 Failure:

A failure is a detected fault. A failure will be declared when the fault cause persisted long enough to consider the ability of an item to perform a required transport function to be terminated. The item

may be considered as failed; a fault has now been detected. See also [ITU-T_G.806]. A failure can be used as a trigger for corrective actions.

3.14 Fault:

A Fault is the inability of a transport function to perform a required action. This does not include an inability due to preventive maintenance, lack of external resources, or planned actions. See also [ITU-T_G.806].

3.15 Layer network:

Layer network is defined in [ITU-T_G.805]. A layer network provides for the transfer of client information and independent operation of the client OAM. A layer network may be described in a service context as follows: one layer network may provide a (transport) service to a higher client layer network and may, in turn, be a client to a lower-layer network. A layer network is a logical construction somewhat independent of arrangement or composition of physical network elements. A particular physical network element may topologically belong to more than one layer network, depending on the actions it takes on the encapsulation associated with the logical layers (e.g., the label stack), and thus could be modeled as multiple logical elements. A layer network may consist of one or more sublayers. For additional explanation of how layer networks relate to the OSI concept of layering, see Appendix I of [ITU-T Y.2611].

3.16 Link:

A physical or logical connection between a pair of Label Switching Routers (LSRs) that are adjacent at the (sub)layer network under consideration. A link may carry zero, one or more LSPs or PWs. A packet entering a link will emerge with the same label stack entry values.

A link as defined in [ITU-T_G.805] is used to describe a fixed relationship between two ports.

3.17 Maintenance Entity (ME):

A Maintenance Entity (ME) can be viewed as the association of two (or more) Maintenance Entity Group End Points (MEPs), that should be configured and managed in order to bound the OAM responsibilities of an OAM flow across a network or sub-network, i.e. a transport path or segment, in the specific layer network that is being monitored

and managed. See also [RFC6371] section 3.1 and [ITU-T G.8113.1], [ITU-T G.8113.2] clause 6.1.

A Maintenance Entity may be defined to monitor and manage bidirectional or unidirectional point-to-point connectivity or point-to-multipoint connectivity in an MPLS-TP layer network.

Therefore, in the context of a MPLS-TP LSP ME or PW ME Label Edge Routers (LERs) and PW Terminating Provider Edges (T-PEs) can be MEPs while LSRs and PW Switching Provider Edges (S-PEs) can be MIPs. In the case of a ME for a Tandem Connection, LSRs and S-PEs can be either MEPs or MIPs.

The following properties apply to all MPLS-TP MEs:

= OAM entities can be nested but not overlapped.

= Each OAM flow is associated to a unique Maintenance Entity.

= OAM packets are subject to the same forwarding treatment as the data traffic, but they are distinct from the data traffic by the Generic Associated Channel Label (GAL).

3.18 Maintenance Entity Group (MEG):

A Maintenance Entity Group is defined, for the purpose of connection monitoring, between a set of connection points within a connection. This set of connection points may be located at the boundary of one administrative domain or a protection domain, or the boundaries of two adjacent administrative domains. The MEG may consist of one or more Maintenance Entities (ME). See also [RFC6371] section 3.1 and [ITU-T G.8113.1], [ITU-T G.8113.2] clause 6.2.

In an MPLS-TP layer network a MEG consists of only one ME.

3.19 Maintenance Entity Group End Point (MEP):

Maintenance Entity Group End Points (MEPs) are the end points of a pre-configured (through the management or control planes) ME. MEPs are responsible for activating and controlling all of the OAM functionality for the ME. A source MEP may initiate an OAM packet to be transferred to its corresponding peer or sink MEP, or to an intermediate MIP that is part of the ME. See also [RFC6371] section 3.3 and [ITU-T G.8113.1], [ITU-T G.8113.2] clause 6.3.

A sink MEP terminates all the OAM packets that it receives corresponding to its ME and does not forward them further along the path.

All OAM packets coming into a source MEP are tunnelled via label stacking and are not processed within the ME as they belong either to the client network layers or to a higher Tandem Connection Monitoring (TCM) level.

A MEP in a tandem connection is not coincident with the termination of the MPLS-TP transport path (LSP or PW), though it can monitor its connectivity (e.g. count packets). A MEP of an MPLS-TP network transport path is coincident with transport path termination and monitors its connectivity (e.g. counts packets).

An MPLS-TP sink MEP can notify a fault condition to its MPLS-TP client layer network.

3.20 Maintenance Entity Group Intermediate Point (MIP):

A Maintenance Entity Group Intermediate Point (MIP) is a point between the two MEPs in an ME and is capable of responding to some OAM packets and forwarding all OAM packets while ensuring fate sharing with data plane packets. A MIP responds only to OAM packets that are sent on the ME it belongs to and that are addressed to the MIP, it does not initiate OAM messages. See also [RFC6371] section 3.4 and [ITU-T G.8113.1], [ITU-T G.8113.2] clause 6.4.

3.21 Management Communication Channel (MCC):

A Communication Channel dedicated for management plane communications.

3.22 Management Communication Network (MCN):

A DCN supporting management plane communication is referred to as a Management Communication Network (MCN).

3.23 Monitoring

Monitoring is applying OAM functionality to verify and to maintain the performance and the quality guarantees of a transport path. There is a need to not only monitor the whole transport path (e.g. LSP or MS-PW), but also arbitrary parts of transport paths. The connection between any two arbitrary points along a transport path is described in one of three ways:
- as a Path Segment Tunnel,

- as a Sub-Path Maintenance Element, or
- as a Tandem Connection.

3.23.1 Path Segment Tunnel (PST):

A path segment is either a segment or a concatenated segment. Path Segment Tunnels (PSTs) are instantiated to provide monitoring of a portion of a set of co-routed transport paths (LSPs or MS-PWs). Path segment tunnels can also be employed to meet the requirement to provide Tandem Connection Monitoring, see Tandem Connection.

3.23.2 Sub-Path Maintenance Element (SPME):

To monitor, protect, and manage a portion (i.e., segment or concatenated segment) of an LSP, a hierarchical LSP [RFC3031] can be instantiated. A hierarchical LSP instantiated for this purpose is called a Sub-Path Maintenance Element (SPME). Note that by definition an SPME does not carry user traffic as a direct client.

An SPME is defined between the edges of the portion of the LSP that needs to be monitored, protected or managed. The SPME forms a MPLS-TP Section that carries the original LSP over this portion of the network as a client. OAM messages can be initiated at the edge of the SPME and sent to the peer edge of the SPME or to a MIP along the SPME. A P router only pushes or pops a label if it is at the end of a SPME. In this mode, it is an LER for the SPME.

3.23.3 Tandem Connection:

A tandem connection is an arbitrary part of a transport path that can be monitored (via OAM) independently from the end-to-end monitoring (OAM). It may be a monitored segment, a monitored concatenated segment or any other monitored ordered sequence of contiguous hops and/or segments (and their interconnecting nodes) of a transport path.

Tandem Connection Monitoring (TCM) for a given path segment of a transport path is implemented by creating a path segment tunnel that has a 1:1 association with the path segment of the transport path that is to be uniquely monitored. This means that the PST used to provide TCM can carry one and only one transport path thus allowing direct correlation between all fault management and performance monitoring information gathered for the PST and the monitored path segment of the end-to-end transport path. The PST is monitored using normal LSP monitoring. See also [RFC6371] section 3.2 and [ITU-T G.8113.1], [ITU-T G.8113.2] clause 6.2.1.

3.24 MPLS Section:

A network segment between two LSRs that are immediately adjacent at the MPLS layer.

3.25 MPLS Transport Profile (MPLS-TP):

The set of MPLS functions used to support packet transport services and network operations.

3.26 MPLS-TP NE:

A network element (NE) that supports MPLS-TP functions.

3.27 MPLS-TP network:

A network in which MPLS-TP NEs are deployed.

3.28 MPLS-TP Recovery:

3.28.1 End-to-end recovery:

MPLS-TP End-to-end recovery refers to the recovery of an entire LSP, from its ingress to its egress node.

3.28.2 Link recovery:

MPLS-TP link recovery refers to the recovery of an individual link (and hence all or a subset of the LSPs routed over the link) between two MPLS-TP nodes. For example, link recovery may be provided by server layer recovery.

3.28.3 Segment recovery:

MPLS-TP Segment recovery refers to the recovery of an LSP segment (i.e., segment and concatenated segment) between two nodes and is used to recover from the failure of one or more links or nodes.

An LSP segment comprises one or more contiguous hops on the path of the LSP. [RFC5654] defines two terms. A "segment" is a single hop along the path of an LSP, while a "concatenated segment" is more than one hop along the path of an LSP.

3.29 MPLS-TP Ring Topology:

In an MPLS-TP ring topology, each LSR is connected to exactly two other LSRs, each via a single point-to-point bidirectional MPLS-TP

capable link. A ring may also be constructed from only two LSRs where there are also exactly two links. Rings may be connected to other LSRs to form a larger network. Traffic originating or terminating outside the ring may be carried over the ring. Client network nodes (such as Customer Edges (CEs)) may be connected directly to an LSR in the ring.

3.29.1 MPLS-TP Logical Ring:

An MPLS-TP logical ring is constructed from a set of LSRs and logical data links (such as MPLS-TP LSP tunnels or MSPL-TP pseudowires) and physical data links that form a ring topology.

3.29.2 MPLS-TP Physical Ring:

An MPLS-TP physical ring is constructed from a set of LSRs and physical data links that form a ring topology.

3.30 OAM flow:

An OAM flow is the set of all OAM packets originating with a specific source MEP that instrument one direction of a MEG (or possibly both in the special case of data plane loopback).

3.31 Operations Support System (OSS):

A system that performs the functions that support processing of information related to operations, administration, maintenance, and provisioning (OAM&P) for the networks, including surveillance and testing functions to support customer access maintenance.

3.32 Path:

See Transport path.

3.33 Protection priority:

Fault conditions (e.g., signal failed), external commands (e.g, forced switch, manual switch) and protection states (e.g., no request) are defined to have a relative priority with respect to each other. Priority is applied to these conditions/command/states locally at each end point and between the two end points.

3.34 Section Layer Network:

A section layer is a server layer (which may be MPLS-TP or a different technology) that provides for the transfer of the section-

layer client information between adjacent nodes in the transport-path layer or transport-service layer. A section layer may provide for aggregation of multiple MPLS-TP clients. Note that [ITU-T_G.805] defines the section layer as one of the two layer networks in a transmission-media layer network. The other layer network is the physical-media layer network.

Section layer networks are concerned with all the functions which provide for the transfer of information between locations in path layer networks.

Physical media layer networks are concerned with the actual fibres, metallic wires or radio frequency channels which support a section layer network.

3.35 Segment:

A link connection as defined in [ITU-T_G.805]. A segment is the part of an LSP that traverses a single link or the part of a PW that traverses a single link (i.e., that connects a pair of adjacent S-PEs and/or T-PEs). See also "Concatenated Segment".

3.36 Server layer:

A server layer is a layer network in which transport paths are used to carry a customer's (individual or bundled) service (may be point-to-point, point-to-multipoint or multipoint-to-multipoint services).

In a client/server relationship (see [ITU-T_G.805]) the server layer network provides a (transport) service to the higher client layer network (usually the layer network under consideration).

3.37 Server MEPs:

A server MEP is a MEP of an ME that is defined in a layer network below the MPLS-TP layer network being referenced. A server MEP coincides with either a MIP or a MEP in the client (MPLS-TP) layer network. See also [RFC6371] section 3.5 and [ITU-T G.8113.1] clause 6.5.

For example, a server MEP can be either:

. A termination point of a physical link (e.g. IEEE 802.3), an SDH VC or OTH ODU for the MPLS-TP Section layer network, defined in [RFC6371] section 3.1.;

. An MPLS-TP Section MEP for MPLS-TP LSPs, defined in [RFC6371] section 3.2.;

. An MPLS-TP LSP MEP for MPLS-TP PWs, defined in [RFC6371] section 3.4.;

. An MPLS-TP TCM MEP for higher-level TCMs, defined in [RFC6371] sections 3.3. and 3.5.

The server MEP can run appropriate OAM functions for fault detection, and notifies a fault indication to the MPLS-TP layer network.

3.38 Signaling Communication Channel (SCC):

A Communication Channel dedicated for control plane communications. The SCC may be used for GMPLS/ASON signaling and/or other control plane messages (e.g., routing messages).

3.39 Signaling Communication Network (SCN):

A DCN supporting control plane communication is referred to as a Signaling Communication Network (SCN).

3.40 Span:

A span is synonymous with a link.

3.41 Sublayer:

Sublayer is defined in [ITU-T_G.805]. The distinction between a layer network and a sublayer is that a sublayer is not directly accessible to clients outside of its encapsulating layer network and offers no direct transport service for a higher layer (client) network.

3.42 Transport Entity:

A "Transport Entity" is a node, link, transport path segment, concatenated transport path segment, or entire transport path.

3.42.1 Working Entity:

A "Working Entity" is a transport entity that carries traffic during normal network operation.

3.42.2 Protection Entity:

A "Protection Entity" is a transport entity that is pre-allocated and used to protect and transport traffic when the working entity fails.

3.42.3 Recovery entity:

A "Recovery Entity" is a transport entity that is used to recover and transport traffic when the working entity fails.

3.43 Transmission media layer:

A layer network, consisting of a section layer network and a physical layer network as defined in [ITU-T_G.805], that provides sections (two-port point-to-point connections) to carry the aggregate of network-transport path or network-service layers on various physical media.

3.44 Transport Network:

A Transport Network provides transmission of traffic between attached client devices by establishing and maintaining point-to-point or point-to-multipoint connections between such devices. A Transport Network is independent of any higher-layer network that may exist between clients, except to the extent required to supply this transmission service. In addition to client traffic, a Transport Network may carry traffic to facilitate its own operation, such as that required to support connection control, network management, and Operations, Administration and Maintenance (OAM) functions.

3.45 Transport path:

A network connection as defined in [ITU-T_G.805]. In an MPLS-TP environment a transport path corresponds to an LSP or a PW.

3.46 Transport path layer:

A (sub)layer network that provides point-to-point or point-to-multipoint transport paths. It provides OAM that is independent of the clients that it is transporting.

3.47 Transport service layer:

A layer network in which transport paths are used to carry a customer's (individual or bundled) service (may be point-to-point, point-to-multipoint or multipoint-to-multipoint services).

3.48 Unidirectional path:

A Unidirectional Path is a path that supports traffic flow in only one direction.

4 Guidance on the Application of this Thesaurus

As discussed in the introduction to this document, this thesaurus is intended to bring the concepts and terms associated with MPLS-TP into the context of the ITU-T's Transport Network architecture. Thus, it should help those familiar with MPLS to see how they may use the features and functions of the Transport Network in order to meet the requirements of MPLS-TP.

5 Management Considerations

The MPLS-TP based network requires management. The MPLS-TP specifications described in [RFC5654], [RFC5860], [RFC5921], [RFC5951], [RFC6371], [RFC6372], [ITU-T G.8110.1] and [ITU-T G.7710], include considerable efforts to provide operator control and monitoring, as well as Operations, Administration and Maintenance (OAM) functionality.

These concepts are, however, out of scope of this document.

6 Security Considerations

Security is a significant requirement of MPLS-TP. See for more information [RFC6941].

However, this informational document is intended only to provide lexicography, and the security concerns are, therefore, out of scope.

7 IANA Considerations

There are no IANA actions resulting from this document.

8 Acknowledgments

The authors would like to thank all members of the teams (the Joint Working Team, the MPLS Interoperability Design Team in IETF and the MPLS-TP Ad Hoc Group in ITU-T) involved in the definition and specification of MPLS Transport Profile. We would in particular like to acknowledge the contributions by Tom Petch to improve the quality of this draft.

9 References

9.1 Normative References

[RFC3031] Rosen, E., Viswanathan, A., and Callon, R., "Multiprotocol Label Switching Architecture", January 2001.

[RFC5654] Niven-Jenkins, B., Brungard, D., Betts, M., et al., "Requirements of an MPLS Transport Profile", September 2009.

[RFC5860] Vigoureux, M., Ward, D., Betts, M., "Requirements for OAM in MPLS Transport Networks", May 2010.

[RFC5921] Bocci, M., Bryant, S., Frost, D, et al., "A Framework for MPLS in Transport Networks", July 2010.

[RFC5951] Lam, K., Gray, E., Mansfield, S., "Network Management Requirements for MPLS-based Transport Networks", September 2010.

[RFC6371] Busi, I., Allan, D., "Operations, Administration, and Maintenance Framework for MPLS-Based Transport Networks", September 2011.

[RFC6372] Sprecher, N., Farrel, A., "MPLS Transport Profile (MPLS-TP) Survivability Framework", September 2011.

For information on the availability of the following documents, please see http://www.itu.int

[ITU-T_G.805] ITU-T Recommendation G.805 (03/2000), "Generic functional architecture of transport networks."

[ITU-T_G.806] ITU-T Recommendation G.806 (03/2006), "Characteristics of transport equipment - Description methodology and generic functionality."

[ITU-T_G.872] ITU-T Recommendation G.872 (11/2001), "Architecture of optical transport networks."

[ITU-T G.7710] ITU-T Recommendation G.7710 (07/2007), "Common equipment management function requirements."

[ITU-T_G.8101] ITU-T Recommendation G.8101/Y.1355 (09/2013), "Terms and definitions for MPLS Transport Profile."

[ITU-T G.8110.1] ITU-T Recommendation G.8110.1/Y.1370.1 (12/2011), "Architecture of the Multi-Protocol Label Switching transport profile layer network."

[ITU-T G.8113.1] ITU-T Recommendation G.8113.1/Y.1372.1 (11/2012), "Operations, Administration and Maintenance mechanism for MPLS-TP in Packet Transport Network (PTN)."

[ITU-T G.8113.2] ITU-T Recommendation G.8113.2/Y.1372.2 (11/2012), "Operations, administration and maintenance mechanisms for MPLS-TP networks using the tools defined for MPLS."

[ITU-T Y.2611] ITU-T Recommendation Y.2611 (12/2006), "High-level architecture of future packet-based networks."

9.2 Informative References

[RFC4397] I. Bryskin, A. Farrel, "A Lexicography for the Interpretation of Generalized Multiprotocol Label Switching (GMPLS) Terminology within the Context of the ITU-T's Automatically Switched Optical Network (ASON) Architecture", February 2006.

[RFC6941] L. Fang, B. Niven-Jenkins, S. Mansfield, R. Graveman, "MPLS Transport Profile (MPLS-TP) Security Framework", April 2013.

Authors' Addresses

Huub van Helvoort (Editor)
Huawei Technologies Co., Ltd.
Email: Huub.van.Helvoort@huawei.com

Loa Andersson (Editor)
Huawei Technologies Co., Ltd.
Email: loa@mail01.huawei.com

Nurit Sprecher (Editor)
Nokia Solutions and Networks
Email: nurit.sprecher@nsn.com

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

S. Kini
Ericsson

January 26, 2011

Hierarchical Labels in the Label Distribution Protocol (LDP)
draft-kini-mpls-ldp-hierarchy-00.txt

Status of this Memo

Distribution of this memo is unlimited.

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/1id-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 July 30, 2011.

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.

Abstract

Label Distribution Protocol (LDP) is used to advertise mappings of Forwarding Equivalence Classes (FECs) to labels. IP prefix FECs are used to setup Label Switched Paths (LSPs) along routed paths. LDP advertises label mappings for IP Prefix FECs that appear as routes in the route table. As the number of FECs in the network increases the number of labels correspondingly increases. During certain failure conditions a large number of label mappings may have to be generated and/or installed in the data plane. In this document we describe an LDP extension to advertise hierarchical label mappings. This helps to reduce the number of label mappings that are downloaded during certain failure conditions and hence improves convergence times.

Table of Contents

1. Introduction

Label Distribution Protocol ([LDP]) is widely deployed protocol in MPLS networks. However it faces limitations in fast convergence especially in scaled scenarios. Some of these problems are described in section 3. A solution to this problem is described in section 4.

2. Conventions used in this document

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

3. Problem Statement

LDP typically allocates a unique label per FEC. Even when the FECs have a common egress LSR, there may be a unique label per FEC. This is typical when the FEC has a unique nexthop on the egress LSR and a FEC lookup on the egress LSR must be avoided. When the path to the egress LSR changes (e.g. due to link/node failures), many FTN and/or ILM entries need to be re-programmed to the data plane. This increases the convergence time.

To deduce the association of the FEC and the egress LSR, a link-state IGP may be used to flood information about the FECs. However, when there are a large number of FECs, IGP stability and convergence are adversely affected. Alternately an additional protocol such as BGP may be used but that increases the operational cost. If targeted LDP is used the number of LDP sessions is of the order of the number of LSRs in the network and this has poor scaling properties.

4. Solution

The solution in this document defines a mechanism to distribute the mapping of a FEC to the corresponding egress LSR (along with the label mapping) using LDP. This label mapping is henceforth referred to as a hierarchical label mapping or H-Label mapping. Using this mapping, an LSP hierarchy is used to transport packets belonging to the FEC. A path to the egress is lower in the hierarchy over which an LSP higher in the hierarchy (specific to the FEC) is tunneled. The label for the LSP higher in the hierarchy is the one allocated by the egress LSR. When the path to an egress LSR changes and results in a nexthop change, only the nexthop corresponding to the path that is lower in the hierarchy needs to be re-programmed in the data plane. This speeds up convergence. Only the FECs to the egress LSR have to be carried in a routing protocol (e.g. link-state IGP), thus reducing the size of the information carried in the routing protocol and results in the faster routing protocol convergence. This also helps

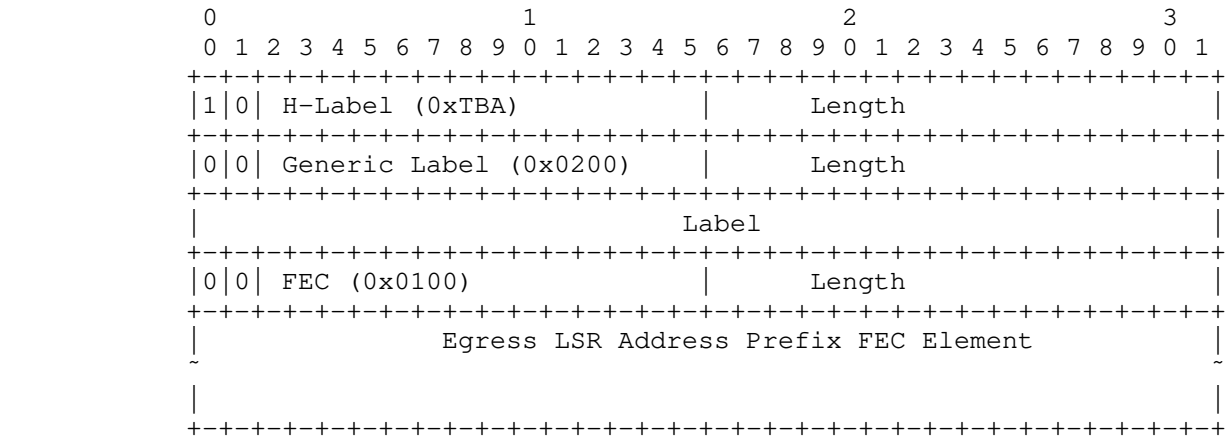
faster routing protocol convergence in cases where the egress LSR goes down.

The new/modified TLVs, messages and procedures to realize this hierarchy are described in subsequent sections.

4.1. TLV Encodings and associated procedures

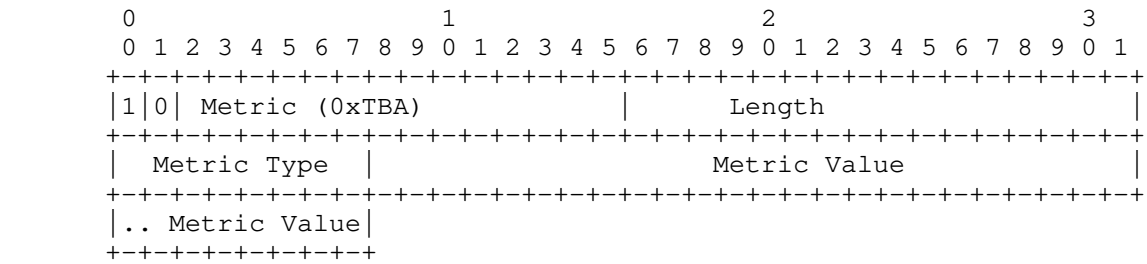
4.1.1. Hierarchical Label (H-Label) TLV

This TLV is a type of Label TLV and can occur in any LDP message that can have a Label TLV as defined in [LDP]. In addition it can occur as an optional parameter in the Label Request message. This TLV consists of two optional sub-TLVs, a Generic Label TLV and a FEC TLV. The FEC TLV MUST have a single Address Prefix FEC Element. This FEC Element is henceforth referred to as Egress LSR Address. When the H-Label TLV occurs in the Label Mapping message it MUST have both TLVs. When the H-Label TLV occurs in the Label Request message it MAY not have any sub-TLV or have just the FEC TLV.



4.1.2. Metric TLV

This TLV is optional in the Label Mapping message that has an H-Label. The Metric is an attribute of the FEC.



Metric Type
1 octet unsigned integer type value. Values in the range 0x00 to 0x0f are defined in this document. Values from 0x10 to 0xff are reserved for future use.

Metric Value
For Metric Types 0x00 to 0x0f this is a 4 byte unsigned integer type value. If the Metric Type is 0x0f then the Metric Value MUST be greater than 0. If a Metric TLV is not present in a label mapping message with an H-Label then the message MUST be treated as having Metric Type of zero and a Metric Value of zero.

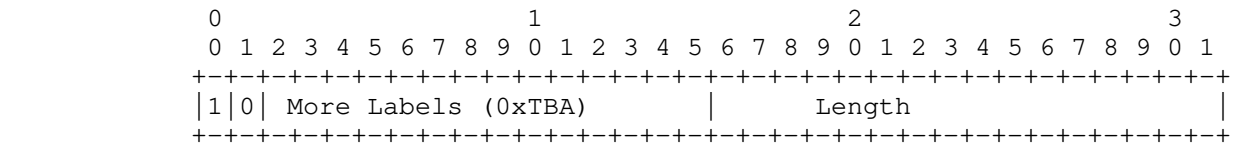
4.1.2.1. Metric TLV procedures

Two Metric TLVs with Metric Type between 0x00 and 0x0f can be compared to determine the lower/higher metric. The comparison procedure is as follows:

1. A metric with Metric Type 'n' is always lower than a metric of Metric Type 'n+1'. This is independent of the Metric Value.
2. If two metrics have the same Metric Type (except if it is 0x0f) then the comparison is made on the value obtained by adding the metric (from the route table) for the route to the Egress Address Prefix FEC Element (from the corresponding H-Label TLV) to the Metric Value.
3. If two metrics have a Metric Type of 0x0f then the comparison is made using only the Metric Value. If the values are the same after comparison, the two metrics are considered equal-cost.

4.1.3. More Label TLV

This TLV is optional in the Label Mapping message that has a Label Request Message ID TLV.



4.1.3.1. More Label TLV procedures

This TLV being present indicates that more label mappings will be sent for that FEC in response to a specific Label Request Message.

4.1.4. H-Label Capability Parameter TLV

This TLV is defined to enable the Hierarchical Label capability. It follows the encoding specified in [LDP-CAP]. There is no Capability Data associated with this TLV. The motivation and behavior when this capability is enabled are all described in this document.

4.2. Extensions to Label distribution and management

4.2.1. Label mapping origination by Egress LSR

An Hierarchical Label capable LSR that is an egress for a FEC due to the nexthop being outside of the label switching network or the FEC elements being reachable by crossing a routing domain boundary SHOULD originate a label mapping with a H-Label. The H-Label MUST have LSR Egress Address as the LSR's own address (typically its loopback address). These mappings SHOULD be advertised to all neighbors that are Hierarchical Label capable. If a neighbor does not have Hierarchical Label capability the LSR MUST advertise labels to it as specified by [LDP].

4.2.2. Label Distribution Control Mode

Both independent and ordered LSP controls are supported when H-Label mappings are advertised.

4.2.2.1. Independent Label Distribution Control

An LSR that is not an egress for a FEC SHOULD advertise H-Label mappings for a FEC with its address (typically loopback) as the Egress LSR Address, if it does not receive H-Label mappings from one of the FEC's nexthops. The LSR MUST additionally advertise a label mapping as described in [LDP] if it has a neighbor that does not have Hierarchical Label Capability.

4.2.2.2. Ordered Label Distribution Control

In this control mode an LSR that receives H-Label mappings from its neighbors, selects those with the lowest cost paths to the FEC. The selection algorithm is described in detail in section 4.3.1.1. These mappings are advertised to its neighbors that are Hierarchical Label capable. Note that routes corresponding to the FECs need not appear in the route table (via IGP) before advertising these label mappings to neighboring LSRs. However, a LSP to the Egress LSR Address must be present. If a neighbor is not Hierarchical Label capable then a label as described in [LDP] is advertised. The FTN and ILM entry creation is described in section 4.2.4.

4.2.3. Label Retention Mode

4.2.3.1. Conservative Label Retention Mode

When using Conservative Label Retention mode, if all the paths/nexthops for the FEC have a common Shared Risk Link Group (SRLG) it is RECOMMENDED that the LSR have as a backup an alternative nexthop that doesn't share an SRLG. This could involve requesting a label from another neighbor. The method to determine the alternate nexthop is outside the scope of this document.

4.2.4. Label Installation

When an H-Label is installed for a FEC in the FTN, packets belonging to the FEC are switched using a hierarchical LSP. The LSP with outer label goes to the egress LSR of the FEC. If some LSRs along the routed path do not support H-Labels, the outer LSP goes till the furthest downstream LSR (that supports H-Labels) along the routed path to the egress LSR before an LSR incapable of H-Label occurs. The outer LSP may even be a TE LSP. The inner LSP identifies the FEC at the egress of the outer LSP.

If a FEC for which a H-Label mapping exists was advertised to a neighbor without the H-Label (due to that neighbor not being capable) then the Incoming Label Map (ILM) entry should be installed with a swap operation to the hierarchical LSP.

If an LSR is purely a transit LSR it SHOULD NOT install any entries in the data plane for label mapping messages with an H-Label.

4.3. Extensions to LDP Messages

This section defines extensions to the LDP Messages and procedures defined in [LDP] and [LDP-IA].

4.3.1. Label Mapping Message

The encoding of the Label Mapping Message has the following modifications:

1. An H-Label TLV can be present instead of a Label TLV. See TLV encoding in section 4.1.1.
2. An optional parameter Metric TLV can occur if an H-Label is present. See TLV encoding in section 4.1.2.

4.3.1.1. Label Mapping Message procedures

Hierarchical Label capable LSRs originate Label Mapping as described in sections 4.2.1. and 4.2.2.1. The Egress LSR Address in the H-Label MUST be an address that belongs to that LSR and has a path from other LSRs. Typically this would be a loopback address for which a LDP label mapping has been advertised. If the metric for the FEC is non-zero then a Metric TLV with appropriate Metric Type and Metric Value is included. An H-Label SHOULD NOT be advertised for addresses that belong to the LSR e.g. loopback addresses.

When an LSR receives Label Mapping Messages for a FEC containing an H-Label, it selects some of these label mappings for advertising to neighbors and installation into the FTN and ILM. The selection algorithm is as follows. Firstly, mappings received from neighbors that are the nexthop for the Egress LSR Address in the corresponding H-Label are chosen. These mappings are advertised to all LDP neighbors. From among these, the mappings with the lowest metric value are chosen using the metric comparison algorithm from section 4.1.2.1. These label mappings are installed in FTN and ILM entries as described in section 4.2.

4.3.2. Label Request Message

The FEC TLV in the request message can contain a Wildcard FEC Element.

4.3.2.1. Label Request Message Procedure

On receiving a Label Request Message, the LSR creates Label Mapping messages for all the Labels for that FEC. If the FEC TLV has a Wildcard FEC Element then all for which label mappings are present are returned in the response. In this case multiple label mapping messages are sent in response. On receiving the LSR additionally checks if it has selected mappings for that FEC according to the procedure in section 4.3.1.1. Those mappings are sent as a response to this request.

4.4. Structuring the FTN

When installing a FTN entry corresponding to a Label Mapping message with a H-Label, a hierarchy must be used. First a push operation using the Label from the Generic Label TLV in the H-Label must be done. This is the LSP for the FEC that is higher in the LSP hierarchy. This must be followed by a push operation of a label for a LSP to the Egress LSR Address. This LSP is lower in the hierarchy. Implementations should handle changes to the nexthops of the LSP to the Egress LSR Address in such a way that the LSPs higher in the hierarchy are quickly updated to realize fast convergence.

5. Mapping metrics of specific IGPs

The 32-bit value in the Metric TLV is sufficient to contain metrics defined for IGPs in [OSPF], [ISIS-TE] and [RIPv2]. The preferences between different routes of an IGP are described within the IGP e.g., [OSPF], [ISIS-2LVL] etc. The Metric Type for the Metric TLV (defined in 4.1.2.) should be chosen in accordance with the IGPs definitions. It should be noted that a metric such as the OSPF Type-2 external metric MUST be mapped to a Metric Type 0xf in the H-Label mapping.

This section describes TLVs that are defined in this document and their associated procedures.

6. Security Considerations

This document does not bring any new security considerations beyond those already described in [LDP].

7. IANA Considerations

The following assignments are required from IANA - TLV Types for Hierarchical Label Capability Parameter, H-Label, Metric and More Label. Recommend next available values 0x604, 0x605, 0x606 and 0x607.

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.

[RIPv2] Malkin, G., "RIP Version 2", RFC 2453, November 1998.

[OSPF] Moy, J., "OSPF Version 2", RFC 2328, April 1998.

[BGP-LBL] Rekhter, Y., "Carrying Label Information in BGP-4", RFC

3107, May 2001.

[LDP] Andersson, L., et al, "LDP Specification", RFC 5036, October 2007.

[LDP-IA] Decraene, B., et al, "LDP Extension for Inter-Area Label Switched Paths (LSPs)", RFC 5283, July 2008.

[ISIS-2LVL] Li, T., et al, "Domain-Wide Prefix Distribution with Two-Level IS-IS", RFC 5302, October 2008.

[LDP-CAP] Thomas, B., et al, "LDP Capabilities", RFC 5561, July 2009.

8.2. Informative References

[MPLS-ARCH] Rosen, E., et al, "Multiprotocol Label Switching Architecture", RFC 3031, January 2001.

[ISIS-TE] Li, T., et al, "IS-IS Extensions for Traffic Engineering", RFC 5307, October 2008.

9. Acknowledgements

The authors would like to thank Joel Halpern and Pramodh D'Souza for their comments.

Authors' Addresses

Sriganesh Kini
Ericsson
300 Holger Way, San Jose, CA 95134
EMail: sriganesh.kini@ericsson.com

Network Working Group
Internet-Draft
Updates: 3031 (if approved)
Intended status: Standards Track
Expires: September 7, 2011

K. Kompella
J. Drake
Juniper Networks
S. Amante
Level 3 Communications, LLC
W. Henderickx
Alcatel-Lucent
L. Yong
Huawei USA

March 6, 2011

The Use of Entropy Labels in MPLS Forwarding
draft-kompella-mpls-entropy-label-02

Abstract

Load balancing is a powerful tool for engineering traffic across a network. This memo suggests ways of improving load balancing across MPLS networks using the concept of "entropy labels". It defines the concept, describes why entropy labels are useful, enumerates properties of entropy labels that allow maximal benefit, and shows how they can be signaled and used for various applications.

Status of this Memo

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

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

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

This Internet-Draft will expire on September 7, 2011.

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.

Table of Contents

1. Introduction

Load balancing, or multi-pathing, is an attempt to balance traffic across a network by allowing the traffic to use multiple paths. Load balancing has several benefits: it eases capacity planning; it can help absorb traffic surges by spreading them across multiple paths; it allows better resilience by offering alternate paths in the event of a link or node failure.

As providers scale their networks, they use several techniques to achieve greater bandwidth between nodes. Two widely used techniques are: Link Aggregation Group (LAG) and Equal-Cost Multi-Path (ECMP). LAG is used to bond together several physical circuits between two adjacent nodes so they appear to higher-layer protocols as a single, higher bandwidth 'virtual' pipe. ECMP is used between two nodes separated by one or more hops, to allow load balancing over several shortest paths in the network. This is typically obtained by arranging IGP metrics such that there are several equal cost paths between source-destination pairs. Both of these techniques may, and often do, co-exist in various parts of a given provider's network, depending on various choices made by the provider.

A very important requirement when load balancing is that packets belonging to a given 'flow' must be mapped to the same path, i.e., the same exact sequence of links across the network. This is to avoid jitter, latency and re-ordering issues for the flow. What constitutes a flow varies considerably. A common example of a flow is a TCP session. Other examples are an L2TP session corresponding to a given broadband user, or traffic within an ATM virtual circuit.

To meet this requirement, a node uses certain fields, termed 'keys', within a packet's header as input to a load balancing function (typically a hash function) that selects the path for all packets in a given flow. The keys chosen for the load balancing function depend on the packet type; a typical set (for IP packets) is the IP source and destination addresses, the protocol type, and (for TCP and UDP traffic) the source and destination port numbers. An overly conservative choice of fields may lead to many flows mapping to the same hash value (and consequently poorer load balancing); an overly aggressive choice may map a flow to multiple values, potentially violating the above requirement.

For MPLS networks, most of the same principles (and benefits) apply. However, finding useful keys in a packet for the purpose of load balancing can be more of a challenge. In many cases, MPLS encapsulation may require fairly deep inspection of packets to find these keys at transit LSRs.

One way to eliminate the need for this deep inspection is to have the ingress LSR of an MPLS Label Switched Path extract the appropriate keys from a given packet, input them to its load balancing function, and place the result in an additional label, termed the 'entropy label', as part of the MPLS label stack it pushes onto that packet.

The packet's MPLS entire label stack can then be used by transit LSRs to perform load balancing, as the entropy label introduces the right level of "entropy" into the label stack.

There are four key reasons why this is beneficial:

1. at the ingress LSR, MPLS encapsulation hasn't yet occurred, so deep inspection is not necessary;

2. the ingress LSR has more context and information about incoming packets than transit LSRs;

3. ingress LSRs usually operate at lower bandwidths than transit LSRs, allowing them to do more work per packet, and

4. transit LSRs do not need to perform deep packet inspection and can load balance effectively using only a packet's MPLS label stack.

This memo describes why entropy labels are needed and defines the properties of entropy labels; in particular how they are generated and received, and the expected behavior of transit LSRs. Finally, it describes in general how signaling works and what needs to be signaled, as well as specifics for the signaling of entropy labels for LDP ([RFC5036]), BGP ([RFC3107], [RFC4364]), and RSVP-TE ([RFC3209]).

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].

The following acronyms are used:

LSR: Label Switching Router;

LER: Label Edge Router;

PE: Provider Edge router;

CE: Customer Edge device; and

FEC: Forwarding Equivalence Class.

The term ingress (or egress) LSR is used interchangeably with ingress (or egress) LER. The term application throughout the text refers to an MPLS application (such as a VPN or VPLS).

A label stack (say of three labels) is denoted by <L1, L2, L3>, where L1 is the "outermost" label and L3 the innermost (closest to the payload). Packet flows are depicted left to right, and signaling is shown right to left (unless otherwise indicated).

The term 'label' is used both for the entire 32-bit label and the 20-bit label field within a label. It should be clear from the context which is meant.

1.2. Motivation

MPLS is very successful generic forwarding substrate that transports several dozen types of protocols, most notably: IP, PWE3, VPLS and IP VPNs. Within each type of protocol, there typically exist several variants, each with a different set of load balancing keys, e.g., for IP: IPv4, IPv6, IPv6 in IPv4, etc.; for PWE3: Ethernet, ATM, Frame-Relay, etc. There are also several different types of Ethernet over PW encapsulation, ATM over PW encapsulation, etc. as well. Finally, given the popularity of MPLS, it is likely that it will continue to be extended to transport new protocols.

Currently, each transit LSR along the path of a given LSP has to try to infer the underlying protocol within an MPLS packet in order to extract appropriate keys for load balancing. Unfortunately, if the transit LSR is unable to infer the MPLS packet's protocol (as is often the case), it will typically use the topmost (or all) MPLS labels in the label stack as keys for the load balancing function. The result may be an extremely inequitable distribution of traffic across equal-cost paths exiting that LSR. This is because MPLS labels are generally fairly coarse-grained forwarding labels that typically describe a next-hop, or provide some of demultiplexing and/or forwarding function, and do not describe the packet's underlying protocol.

On the other hand, an ingress LSR (e.g., a PE router) has detailed knowledge of an packet's contents, typically through a priori configuration of the encapsulation(s) that are expected at a given PE-CE interface, (e.g., IPv4, IPv6, VPLS, etc.). They also have more flexible forwarding hardware. PE routers need this information and these capabilities to:

a) apply the required services for the CE;

b) discern the packet's CoS forwarding treatment;

c) apply filters to forward or block traffic to/from the CE;

d) to forward routing/control traffic to an onboard management processor; and,

e) load-balance the traffic on its uplinks to transit LSRs (e.g., P routers).

By knowing the expected encapsulation types, an ingress LSR router can apply a more specific set of payload parsing routines to extract the keys appropriate for a given protocol. This allows for significantly improved accuracy in determining the appropriate load balancing behavior for each protocol.

If the ingress LSR were to capture the flow information so gathered in a convenient form for downstream transit LSRs, transit LSRs could remain completely oblivious to the contents of each MPLS packet, and use only the captured flow information to perform load balancing. In particular, there will be no reason to duplicate an ingress LSR's complex packet/payload parsing functionality in a transit LSR. This will result in less complex transit LSRs, enabling them to more easily scale to higher forwarding rates, larger port density, lower power consumption, etc. The idea in this memo is to capture this flow information as a label, the so-called entropy label.

Ingress LSRs can also adapt more readily to new protocols and extract the appropriate keys to use for load balancing packets of those protocols. This means that deploying new protocols or services in edge devices requires fewer concommitant changes in the core, resulting in higher edge service velocity and at the same time more stable core networks.

2. Approaches

There are two main approaches to encoding load balancing information in the label stack. The first allocates multiple labels for a particular Forwarding Equivalance Class (FEC). These labels are equivalent in terms of forwarding semantics, but having multiple labels allows flexibility in assigning labels to flows belonging to the same FEC. This approach has the advantage that the label stack has the same depth whether or not one uses label-based load balancing; and so, consequently, there is no change to forwarding operations on transit and egress LSRs. However, it has a major

drawback in that there is a significant increase in both signaling and forwarding state.

The other approach encodes the load balancing information as an additional label in the label stack, thus increasing the depth of the label stack by one. With this approach, there is minimal change to signaling state for a FEC; also, there is no change in forwarding operations in transit LSRs, and no increase of forwarding state in any LSR. The only purpose of the additional label is to increase the entropy in the label stack, so this is called an "entropy label". This memo focuses solely on this approach.

3. Entropy Labels

An entropy label (as used here) is a label:

1. that is not used for forwarding;

2. that is not signaled; and

3. whose only purpose in the label stack is to provide 'entropy' to improve load balancing.

Entropy labels are generated by an ingress LSR, based entirely on load balancing information. However, they MUST NOT have values in the reserved label space (0-15). Entropy labels MUST be at the bottom of the label stack, and thus the 'Bottom of Stack' (S) bit ([RFC3032]) in the label should be set. To ensure that they are not used inadvertently for forwarding, entropy labels SHOULD have a TTL of 0.

Since entropy labels are generated by an ingress LSR, an egress LSR MUST be able to tell unambiguously that a given label is an entropy label. If any ambiguity is possible, the label above the entropy label MUST be an 'entropy label indicator' (ELI), which indicates that the following Label is an entropy label. An ELI is typically signaled by an egress LSR and is added to the MPLS label stack along with an entropy label by an ingress LSR. For many applications, the use of entropy labels is unambiguous, and an ELI is not needed. If used, an ELI MUST have S = 0 and SHOULD have a TTL of 0.

Applications for MPLS entropy labels include pseudowires ([RFC4447]), Layer 3 VPNs ([RFC4364]), VPLS ([RFC4761], [RFC4762]) and Tunnel LSPs carrying, say, IP traffic. [I-D.ietf-pwe3-fat-pw] explains how entropy labels can be used for RFC 4447-style pseudowires, and thus is complementary to this memo, which focuses on several other applications of entropy labels.

4. Data Plane Processing of Entropy Labels

4.1. Ingress LSR

Suppose that for a particular application (or service or FEC), an ingress LSR X is to push label stack <TL, AL>, where TL is the 'tunnel label' and AL is the 'application label'. (Note the use of the convention for label stacks described in Section 1.1. The use of a two-label stack is just for illustrative purposes.) Suppose furthermore that the egress LSR Y has told X that it is capable of processing entropy labels for this application. If X can insert entropy labels, it may use a label stack of <TL, AL, EL> for this application, where EL is the entropy label.

When a packet for this application arrives at X, X does the following:

1. X identifies the application to which the packet belongs, identifies the egress LSR as Y, and thereby picks the outgoing label stack <TL, AL> to push onto the packet to send to Y;

2. X determines which keys that it will use for load balancing;

3. X, having kept state that Y can process entropy labels for this application, generates an entropy label EL (based on the output of the load balancing function), and

4. X pushes <TL, AL, EL> on to the packet before forwarding it to the next LSR on its way to Y.

EL is a 'regular' 32-bit label whose S bit MUST be 1 and whose TTL field SHOULD be 0. The load balancing information is encoded in the 20-bit label field. If X is told (via signaling) that it must use an entropy label indicator with label value E, then X instead pushes <TL, AL, ELI, EL> onto the packet, where ELI is a label whose S bit MUST be 0, whose TTL SHOULD be 0, and whose 20-bit label field MUST be E. The CoS fields for EL and ELI can be set to any values.

Note that ingress LSR X MUST NOT include an entropy label unless the egress LSR Y for this application has indicated that it is ready to receive entropy labels. Furthermore, if Y has signaled that an ELI is needed, then X MUST include the ELI before the entropy label.

Note that the signaling and use of entropy labels in one direction (signaling from Y to X, and data path from X to Y) has no bearing on the behavior in the opposite direction (signaling from X to Y, and data path from Y to X).

4.2. Transit LSR

Transit LSRs have virtually no change in forwarding behavior. For load balancing, transit LSRs SHOULD use the whole label stack as keys for the load balancing function. Transit LSRs MAY choose to look beyond the label stack for further keys; however, if entropy labels are being used, this may not be very useful. Looking beyond the label stack may be the simplest approach in an environment where some ingress LSRs use entropy labels and others don't, or for backward compatibility. Thus, other than using the full label stack as input to the load balancing function, transit LSRs are almost unaffected by the use of entropy labels.

4.3. Egress LSR

If egresss LSR Y signals that it is capable of processing entropy labels without an ELI for an application, then when Y receives a packet with the application label, then Y looks to see if the S bit is set. If so, Y applies its usual processing rules to the packet, including popping the application label. If the S bit is not set, Y assumes that the label below the application label is an entropy label and pops both the application label and the entropy label. Y SHOULD ensure that the entropy label has its S bit set. Y then processes the packet as usual. Implementations may choose the order in which they apply these operations, but the net result should be as specified.

If Y signals that it is capable of processing entropy labels but that an ELI is necessary for a given application, then when Y receives a packet with the application label, Y processes the application label as usual, then pops it. Y then checks whether the S bit on the application label is set. If not, Y looks to see if the label below the application label is the ELI. If so, Y further pops both the ELI and the label below (which should be the entropy label). Y SHOULD ensure that the ELI has its S bit unset, and that the entropy label has its S bit set. If the S bit of the application label is set, or the label below is not the ELI, Y processes the packet as usual (there is no entropy label).

5. Signaling for Entropy Labels

An egress LSR Y may signal to ingress LSR(s) its ability to process entropy labels on a per-application (or per-FEC) basis. As part of this signaling, Y also signals the ELI to use, if any.

In cases where an application label is used and must be the bottommost label in the label stack, Y MAY signal that no ELI is

needed for that application.

In cases where no application label exists, or where the application label may not be the bottommost label in the label stack, Y MUST signal a valid ELI to be used in conjunction with the entropy label for this FEC. In this case, an ingress LSR will either not add an entropy label, or push the ELI before the entropy label. This makes the use or non-use of an entropy label by the ingress LSR unambiguous. Valid ELI label values are strictly greater than 15.

It should be noted that egress LSR Y may use the same ELI value for all applications for which an ELI is needed. The ELI MUST be a label that does not conflict with any other labels that Y has advertised to other LSRs for other applications. Furthermore, it should be noted that the ability to process entropy labels (and the corresponding ELI) may be asymmetric: an LSR X may be willing to process entropy labels, whereas LSR Y may not be willing to process entropy labels. The signaling extensions below allow for this asymmetry.

For an illustration of signaling and forwarding with entropy labels, see Figure 9.

5.1. LDP Signaling

When using LDP for signaling tunnel labels ([RFC5036]), a Label Mapping Message sub-TLV (Entropy Label sub-TLV) is used to signal an egress LSR's ability to process entropy labels.

The presence of the Entropy Label sub-TLV in the Label Mapping Message indicates to ingress LSRs that the egress LSR can process an entropy label. In addition, the Entropy Label sub-TLV contains a label value for the ELI. If the ELI is zero, this indicates the egress doesn't need an ELI for the signaled application; if not, the egress requires the given ELI with entropy labels. An example where an ELI is needed is when the signaled application is an LSP that can carry IP traffic.

The structure of the Entropy Label sub-TLV is shown below.

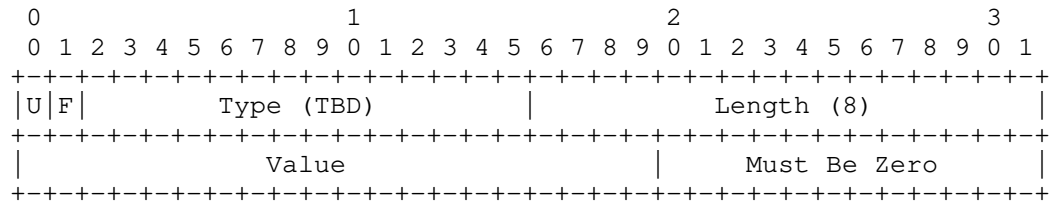


Figure 1: Entropy Label sub-TLV

where:

U: Unknown bit. This bit MUST be set to 1. If the Entropy Label sub-TLV is not understood, then the TLV is not known to the receiver and MUST be ignored.

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

Type: sub-TLV Type field, as specified by IANA.

Length: sub-TLV Length field. This field specifies the total length in octets of the Entropy Label sub-TLV.

Value: value of the Entropy Label Indicator Label.

5.2. BGP Signaling

When BGP [RFC4271] is used for distributing Network Layer Reachability Information (NLRI) as described in, for example, [RFC3107], [RFC4364] and [RFC4761], the BGP UPDATE message may include the Entropy Label attribute. This is an optional, transitive BGP attribute of type TBD. The inclusion of this attribute with an NLRI indicates that the advertising BGP router can process entropy labels as an egress LSR for that NLRI. If the attribute length is less than three octets, this indicates that the egress doesn't need an ELI for the signaled application. If the attribute length is at least three octets, the first three octets encode an ELI label value as the high order 20 bits; the egress requires this ELI with entropy labels. An example where an ELI is needed is when the NLRI contains unlabeled IP prefixes.

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

A1: S sets the BGP NEXT_HOP attribute to itself; AND

A2: S can process entropy labels for the given application.

If both A1 and A2 are true, and S needs an ELI to recognize entropy labels, then S MUST include the ELI label value as part of the Entropy Label attribute. An UPDATE SHOULD contain at most one Entropy Label attribute.

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

B1: T does not change the NEXT_HOP attribute; OR

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

An example of the use of B1 is Route Reflectors; an example of the use of B2 is illustrated in Section 9.3.1.2.

However, if T changes the NEXT_HOP attribute for U and in the data plane pops the entire label stack to process the payload, T MUST remove ELA. T MAY include a new Entropy Label attribute ELA' for UPDATE U' if both of the following are true:

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

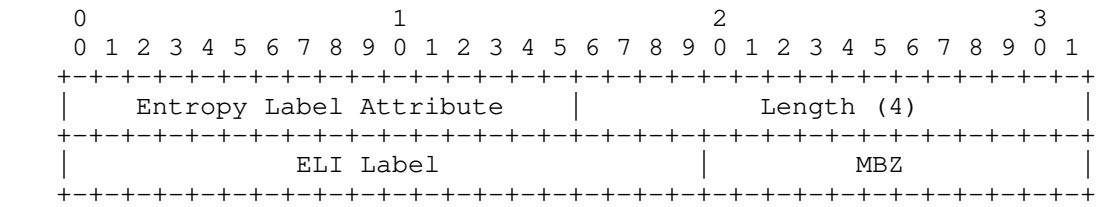
C2: T can process entropy labels for the given application.

Again, if both C1 and C2 are true, and T needs an ELI to recognize entropy labels, then T MUST include the ELI label value as part of the Entropy Label attribute.

5.3. RSVP-TE Signaling

Entropy Label support is signaled in RSVP-TE [RFC3209] using an Entropy Label Attribute TLV (Type TBD) of the LSP_ATTRIBUTES object [RFC5420]. The presence of this attribute indicates that the signaler (the egress in the downstream direction using Resv messages; the ingress in the upstream direction using Path messages) can process entropy labels. The Entropy Label Attribute contains a value for the ELI. If the ELI is zero, this indicates that the signaler doesn't need an ELI for this application; if not, then the signaler requires the given ELI with entropy labels. An example where an ELI is needed is when the signaled LSP can carry IP traffic.

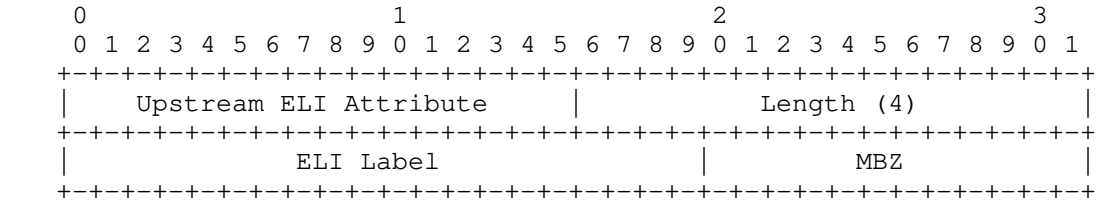
The format of the Entropy Label Attribute is as follows:



An egress LSR includes the Entropy Label Attribute in a Resv message to indicate that it can process entropy labels in the downstream direction of the signaled LSP.

An ingress LSR includes the Entropy Label Attribute in a Path message for a bi-directional LSP to indicate that it can process entropy labels in the upstream direction of the signaled LSP. If the signaled LSP is not bidirectional, the Entropy Label Attribute SHOULD NOT be included in the Path message, and egress LSR(s) SHOULD ignore the attribute, if any.

As described in Section 8, there is also the need to distribute an ELI from the ingress (upstream label allocation). In the case of RSVP-TE, this is accomplished using the Upstream ELI Attribute TLV of the LSP_ATTRIBUTES object, as shown below:



6. Operations, Administration, and Maintenance (OAM) and Entropy Labels

Generally OAM comprises a set of functions operating in the data plane to allow a network operator to monitor its network infrastructure and to implement mechanisms in order to enhance the general behavior and the level of performance of its network, e.g., the efficient and automatic detection, localization, diagnosis and handling of defects.

Currently defined OAM mechanisms for MPLS include LSP Ping/Traceroute [RFC4379] and Bidirectional Failure Detection (BFD) for MPLS [RFC5884]. The latter provides connectivity verification between the endpoints of an LSP, and recommends establishing a separate BFD session for every path between the endpoints.

The LSP traceroute procedures of [RFC4379] allow an ingress LSR to obtain label ranges that can be used to send packets on every path to the egress LSR. It works by having ingress LSR sequentially ask the transit LSRs along a particular path to a given egress LSR to return a label range such that the inclusion of a label in that range in a packet will cause the replying transit LSR to send that packet out the egress interface for that path. The ingress provides the label range returned by transit LSR N to transit LSR N + 1, which returns a label range which is less than or equal in span to the range provided to it. This process iterates until the penultimate transit LSR replies to the ingress LSR with a label range that is acceptable to it and to all LSRs along path preceding it for forwarding a packet along the path.

However, the LSP traceroute procedures do not specify where in the label stack the value from the label range is to be placed, whether deep packet inspection is allowed and if so, which keys and key values are to be used.

This memo updates LSP traceroute by specifying that the value from the label range is to be placed in the entropy label. Deep packet inspection is thus not necessary, although an LSR may use it, provided it do so consistently, i.e., if the label range to go to a given downstream LSR is computed with deep packet inspection, then the data path should use the same approach and the same keys.

In order to have a BFD session on a given path, a value from the label range for that path should be used as the EL value for BFD packets sent on that path.

As part of the MPLS-TP work, an in-band OAM channel is defined in [RFC5586]. Packets sent in this channel are identified with a reserved label, the Generic Associated Channel Label (GAL) placed at the bottom of the MPLS label stack. In order to use the inband OAM channel with entropy labels, this memo relaxes the restriction that the GAL must be at the bottom of the MPLS label stack. Rather, the GAL is placed in the MPLS label stack above the entropy label so that it effectively functions as an application label.

7. MPLS-TP and Entropy Labels

Since MPLS-TP does not use ECMP, entropy labels are not applicable to an MPLS-TP deployment.

8. Point-to-Multipoint LSPs and Entropy Labels

Point-to-Multipoint (P2MP) LSPs [RFC4875] typically do not use ECMP for load balancing, as the combination of replication and multipathing can lead to duplicate traffic delivery. However, P2MP LSPs can traverse Bundled Links [RFC4201] and LAGs. In both these cases, load balancing is useful, and hence entropy labels can be of some value for P2MP LSPs.

There are two potential complications with the use of entropy labels in the context of P2MP LSPs, both a consequence of the fact that the entire label stack below the P2MP label must be the same for all egress LSRs. First, all egress LSRs must be willing to receive entropy labels; if even one egress LSR is not willing, then entropy labels MUST NOT be used for this P2MP LSP. Second, if an ELI is required, all egress LSRs must agree to the same value of ELI. This can be achieved by upstream allocation of the ELI; in particular, for RSVP-TE P2MP LSPs, the ingress LSR distributes the ELI value using the Upstream ELI Attribute TLV of the LSP_ATTRIBUTES object, defined in Section 5.3.

With regard to the first issue, the ingress LSR MUST keep track of the ability of each egress LSR to process entropy labels, especially since the set of egress LSRs of a given P2MP LSP may change over time. Whenever an existing egress LSR leaves, or a new egress LSR joins the P2MP LSP, the ingress MUST re-evaluate whether or not to include entropy labels for the P2MP LSP.

In some cases, it may be feasible to deploy two P2MP LSPs, one to entropy label capable egress LSRs, and the other to the remaining egress LSRs. However, this requires more state in the network, more bandwidth, and more operational overhead (tracking EL-capable LSRs, and provisioning P2MP LSPs accordingly). Furthermore, this approach may not work for some applications (such mVPNs and VPLS) which automatically create and/or use P2MP LSPs for their multicast requirements.

9. Entropy Labels and Applications

This section describes the usage of entropy labels in various scenarios with different applications.

9.1. Tunnels

Tunnel LSPs, signaled with either LDP or RSVP-TE, typically carry other MPLS applications such as VPNs or pseudowires. This being the case, if the egress LSR of a tunnel LSP is willing to process entropy

labels, it would signal the need for an Entropy Label Indicator to distinguish between entropy labels and other application labels.

In the figures below, the following convention is used to depict information signaled between X and Y:

```
                    X ---------- ... ---------- Y
               app:   <--- [label L, ELI value]
```

This means Y signals to X label L for application app. The ELI value can be one of:

-: meaning entropy labels are NOT accepted;

0: meaning entropy labels are accepted, no ELI is needed; or

E: entropy labels are accepted, ELI label E is required.

The following illustrates a simple intra-AS tunnel LSP.

```
                X -------- A --- ... --- B -------- Y
   tunnel LSP L:   [TL,  E] <---  ...  <--- [TL0, E]

   IP pkt:         push <TL, E, EL> --------------->
```

Figure 2: Tunnel LSPs and Entropy Labels

Tunnel LSPs may cross Autonomous System (AS) boundaries, usually using BGP ([RFC3107]). In this case, the AS Border Routers (ASBRs) MAY simply propagate the egress LSR's ability to process entropy labels, or they MAY declare that entropy labels may not be used. If an ASBR (say A2 below) chooses to propagate the egress LSR Y's ability to process entropy labels, A2 MUST also propagate Y's choice of ELI.

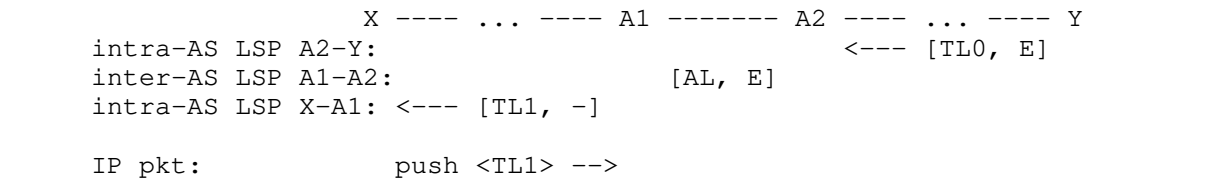
```
              X ---- ... ---- A1 ------- A2 ---- ... ---- Y
  intra-AS LSP A2-Y:                          <--- [TL0, E]
  inter-AS LSP A1-A2:              [AL, E]
  intra-AS LSP X-A1: <--- [TL1, E]

  IP pkt:          push <TL1, E, EL>
```

Here, ASBR A2 chooses to propagate Y's ability to process entropy labels, by "translating" Y's signaling of entropy label capability (say using LDP) to BGP; and A1 translate A2's BGP signaling to (say) RSVP-TE. The end-to-end tunnel (X to Y) will have entropy labels if

X chooses to insert them.

Figure 3: Inter-AS Tunnel LSP with Entropy Labels



Here, ASBR A1 decided that entropy labels are not to be used; thus, the end-to-end tunnel cannot have entropy labels, even though both X and Y may be capable of inserting and processing entropy labels.

Figure 4: Inter-AS Tunnel LSP with no Entropy Labels

9.2. LDP Pseudowires

[I-D.ietf-pwe3-fat-pw] describes the signaling and use of entropy labels in the context of RFC 4447 pseudowires, so this will not be described further here.

[RFC4762] specifies the use of LDP for signaling VPLS pseudowires. An egress VPLS PE that can process entropy labels can indicate this by adding the Entropy Label sub-TLV in the LDP message it sends to other PEs. An ELI is not required. An ingress PE must maintain state per egress PE as to whether it can process entropy labels.

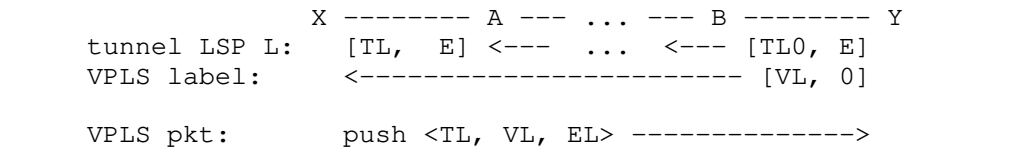


Figure 5: Entropy Labels with LDP VPLS

Note that although the underlying tunnel LSP signaling indicated the need for an ELI, VPLS packets don't need an ELI, and thus the label stack pushed by X do not have one.

[RFC4762] also describes the notion of "hierarchical VPLS" (H-VPLS). In H-VPLS, 'hub PEs' remove the label stack and process VPLS packets; thus, they must make their own decisions on the use of entropy labels, independent of other hub PEs or spoke PEs with which they exchange signaling. In the example below, spoke PEs X and Y and hub

PE B can process entropy labels, but hub PE A cannot.

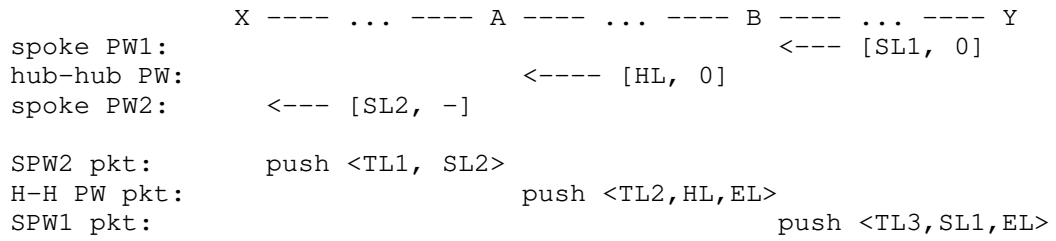


Figure 6: Entropy Labels with H-VPLS

9.3. BGP Applications

Section 9.1 described a BGP application for the creation of inter-AS tunnel LSPs. This section describes two other BGP applications, IP VPNs ([RFC4364]) and BGP VPLS ([RFC4761]). An egress PE for either of these applications indicates its ability to process entropy labels by adding the Entropy Label attribute to its BGP UPDATE message. Again, ingress PEs must maintain per-egress PE state regarding its ability to process entropy labels. In this section, both of these applications will be referred to as VPNs.

In the intra-AS case, PEs signal application labels and entropy label capability to each other, either directly, or via Route Reflectors (RRs). If RRs are used, they must not change the BGP NEXT_HOP attribute in the UPDATE messages; furthermore, they can simply pass on the Entropy Label attribute as is.

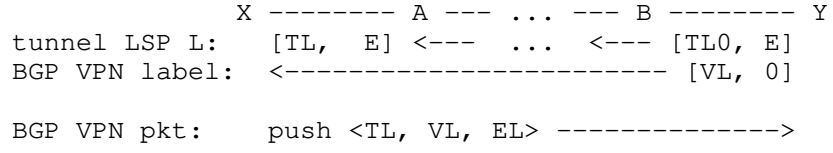


Figure 7: Entropy Labels with Intra-AS BGP apps

For BGP VPLS, the application label is at the bottom of stack, so no ELI is needed. For BGP IP VPNs, the application label is usually at the bottom of stack, so again no ELI is needed. However, in the case of Carrier's Carrier (CsC) VPNs, the BGP VPN label may not be at the bottom of stack. In this case, an ELI is necessary for CsC VPN packets with entropy labels to distinguish them from nested VPN packets. In the example below, the nested VPN signaling is not shown; the egress PE for the nested VPN (not shown) must signal

whether or not it can process egress labels, and the ingress nested VPN PE may insert an entropy label if so.

Three cases are shown: a plain BGP VPN packet, a CsC VPN packet originating from X, and a transit nested VPN packet originating from a nested VPN ingress PE (conceptually to the left of X). It is assumed that the nested VPN packet arrives at X with label stack <ZL, CVL> where ZL is the tunnel label (to be swapped with <TL, CL>) and CVL is the nested VPN label. Note that Y can use the same ELI for the tunnel LSP and the CsC VPN (and any other application that needs an ELI).

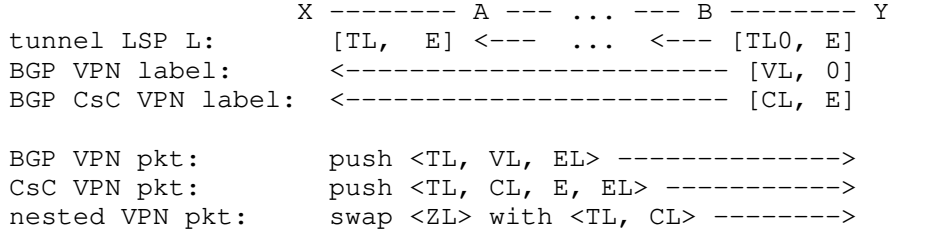


Figure 8: Entropy Labels with CoC VPN

9.3.1. Inter-AS BGP VPNs

There are three commonly used options for inter-AS IP VPNs and BGP VPLS, known informally as "Option A", "Option B" and "Option C". This section describes how entropy labels can be used in these options.

9.3.1.1. Option A Inter-AS VPNs

In option A, an ASBR pops the full label stack of a VPN packet exiting an AS, processes the payload header (IP or Ethernet), and forwards the packet natively (i.e., as IP or Ethernet, but not as MPLS) to the peer ASBR. Thus, entropy label signaling and insertion are completely local to each AS. The inter-AS paths do not use entropy labels, as they do not use a label stack.

9.3.1.2. Option B Inter-AS VPNs

The ASBRs in option B inter-AS VPNs have a choice (usually determined by configuration) of whether to just swap labels (from within the AS to the neighbor AS or vice versa), or to pop the full label stack and process the packet natively. This choice occurs at each ASBR in each direction. In the case of native packet processing at an ASBR, entropy label signaling and insertion is local to each AS and to the

inter-AS paths (which, unlike option A, do have labeled packets).

In the case of simple label swapping at an ASBR, the ASBR can propagate received entropy label signaling onward. That is, if a PE signals to its ASBR that it can process entropy labels (via an Entropy Label attribute), the ASBR can propagate that attribute to its peer ASBR; if a peer ASBR signals that it can process entropy labels, the ASBR can propagate that to all PEs within its AS). Note that this is the case even though ASBRs change the BGP NEXT_HOP attribute to "self", because of clause B2 in Section 5.2.

9.3.1.3. Option C Inter-AS VPNs

In Option C inter-AS VPNs, the ASBRs are not involved in signaling; they do not have VPN state; they simply swap labels of inter-AS tunnels. Signaling is PE to PE, usually via Route Reflectors; however, if RRs are used, the RRs do not change the BGP NEXT_HOP attribute. Thus, entropy label signaling and insertion are on a PE-pair basis, and the intermediate routers, ASBRs and RRs do not play a role.

9.4. Multiple Applications

It has been mentioned earlier that an ingress PE must keep state per egress PE with regard to its ability to process entropy labels. An ingress PE must also keep state per application, as entropy label processing must be based on the application context in which a packet is received (and of course, the corresponding entropy label signaling).

In the example below, an egress LSR Y signals a tunnel LSP L, and is prepared to receive entropy labels on L, but requires an ELI. Furthermore, Y signals two pseudowires PW1 and PW2 with labels PL1 and PL2, respectively, and indicates that it can receive entropy labels for both pseudowires without the need of an ELI; and finally, Y signals a L3 VPN with label VL, but Y does not indicate that it can receive entropy labels for the L3 VPN. Ingress LSR X chooses to send native IP packets to Y over L with entropy labels, thus X must include the given ELI (yielding a label stack of <TL, ELI, EL>). X chooses to add entropy labels on PW1 packets to Y, with a label stack of <TL, PL1, EL>, but chooses not to do so for PW2 packets. X must not send entropy labels on L3 VPN packets to Y, i.e., the label stack must be <TL, VL>.

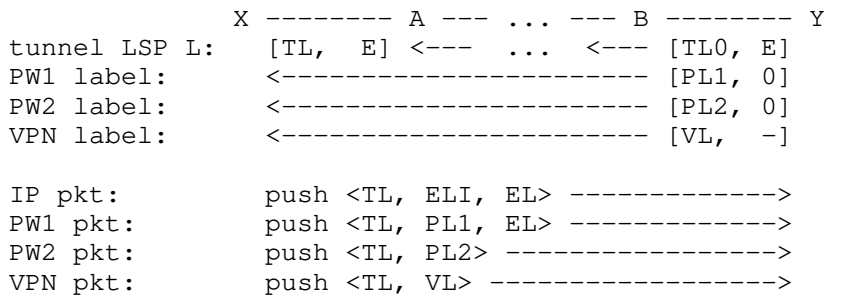


Figure 9: Entropy Labels for Multiple Applications

10. Security Considerations

This document describes advertisement of the capability to support receipt of entropy-labels and an Entropy Label Indicator that an ingress LSR may apply to MPLS packets in order to allow transit LSRs to attain better load-balancing across LAG and/or ECMP paths in the network.

This document does not introduce new security vulnerabilities to LDP. Please refer to the Security Considerations section of LDP ([RFC5036]) for security mechanisms applicable to LDP.

Given that there is no end-user control over the values used for entropy labels, there is little risk of Entropy Label forgery which could cause uneven load-balancing in the network.

If Entropy Label Capability is not signaled from an egress PE to an ingress PE, due to, for example, malicious configuration activity on the egress PE, then the PE's will fall back to not using entropy labels for load-balancing traffic over LAG or ECMP paths which, in some cases, in no worse than the behavior observed in current production networks. That said, operators are recommended to monitor changes to PE configurations and, more importantly, the fairness of load distribution over equal-cost LAG or ECMP paths. If the fairness of load distribution over a set of paths changes that could indicate a misconfiguration, bug or other non-optimal behavior on their PE's and they should take corrective action.

Given that most applications already signal an Application Label, e.g.: IPVPNs, LDP VPLS, BGP VPLS, whose Bottom of Stack bit is being re-used to signal entropy label capability, there is little to no additional risk that traffic could be misdirected into an inappropriate IPVPN VRF or VPLS VSI at the egress PE.

In the context of downstream-signaled entropy labels that require the use of an Entropy Label Indicator (ELI), there should be little to no additional risk because the egress PE is solely responsible for allocating an ELI value and ensuring that ELI label value DOES NOT conflict with other MPLS labels it has previously allocated. On the other hand, for upstream-signaled entropy labels, e.g.: RSVP-TE point-to-point or point-to-multipoint LSP's or Multicast LDP (mLDP) point-to-multipoint or multipoint-to-multipoint LSP's, there is a risk that the head-end MPLS LER may choose an ELI value that is already in use by a downstream LSR or LER. In this case, it is the responsibility of the downstream LSR or LER to ensure that it MUST NOT accept signaling for an ELI value that conflicts with MPLS label(s) that are already in use.

11. IANA Considerations

11.1. LDP Entropy Label TLV

IANA is requested to allocate the next available value from the IETF Consensus range in the LDP TLV Type Name Space Registry as the "Entropy Label TLV".

11.2. BGP Entropy Label Attribute

IANA is requested to allocate the next available Path Attribute Type Code from the "BGP Path Attributes" registry as the "BGP Entropy Label Attribute".

11.3. Attribute Flags for LSP_Attributes Object

IANA is requested to allocate a new bit from the "Attribute Flags" sub-registry of the "RSVP TE Parameters" registry.

Bit No	Name	Attribute Flags Path	Attribute Flags Resv	RRO
TBD	Entropy Label LSP	Yes	Yes	No

11.4. Attributes TLV for LSP_Attributes Object

IANA is requested to allocate the next available value from the "Attributes TLV" sub-registry of the "RSVP TE Parameters" registry.

12. Acknowledgments

We wish to thank Ulrich Drafz for his contributions, as well as the entire 'hash label' team for their valuable comments and discussion.

13. References

13.1. 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.

[RFC3107] Rekhter, Y. and E. Rosen, "Carrying Label Information in BGP-4", RFC 3107, May 2001.

[RFC3209] Awduche, D., Berger, L., Gan, D., Li, T., Srinivasan, V., and G. Swallow, "RSVP-TE: Extensions to RSVP for LSP Tunnels", RFC 3209, December 2001.

[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.

13.2. Informative References

[I-D.ietf-pwe3-fat-pw] Bryant, S., Filsfils, C., Drafz, U., Kompella, V., Regan, J., and S. Amante, "Flow Aware Transport of Pseudowires over an MPLS PSN", draft-ietf-pwe3-fat-pw-05 (work in progress), October 2010.

[RFC4201] Kompella, K., Rekhter, Y., and L. Berger, "Link Bundling in MPLS Traffic Engineering (TE)", RFC 4201, October 2005.

[RFC4271] Rekhter, Y., Li, T., and S. Hares, "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, January 2006.

[RFC4364] Rosen, E. and Y. Rekhter, "BGP/MPLS IP Virtual Private Networks (VPNs)", RFC 4364, February 2006.

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

February 2006.

[RFC4447] Martini, L., Rosen, E., El-Aawar, N., Smith, T., and G. Heron, "Pseudowire Setup and Maintenance Using the Label Distribution Protocol (LDP)", RFC 4447, April 2006.

[RFC4761] Kompella, K. and Y. Rekhter, "Virtual Private LAN Service (VPLS) Using BGP for Auto-Discovery and Signaling", RFC 4761, January 2007.

[RFC4762] Lasserre, M. and V. Kompella, "Virtual Private LAN Service (VPLS) Using Label Distribution Protocol (LDP) Signaling", RFC 4762, January 2007.

[RFC4875] Aggarwal, R., Papadimitriou, D., and S. Yasukawa, "Extensions to Resource Reservation Protocol - Traffic Engineering (RSVP-TE) for Point-to-Multipoint TE Label Switched Paths (LSPs)", RFC 4875, May 2007.

[RFC5036] Andersson, L., Minei, I., and B. Thomas, "LDP Specification", RFC 5036, October 2007.

[RFC5586] Bocci, M., Vigoureux, M., and S. Bryant, "MPLS Generic Associated Channel", RFC 5586, June 2009.

[RFC5884] Aggarwal, R., Kompella, K., Nadeau, T., and G. Swallow, "Bidirectional Forwarding Detection (BFD) for MPLS Label Switched Paths (LSPs)", RFC 5884, June 2010.

Appendix A. Applicability of LDP Entropy Label sub-TLV

In the case of unlabeled IPv4 (Internet) traffic, the Best Current Practice is for an egress LSR to propagate eBGP learned routes within a SP's Autonomous System after resetting the BGP next-hop attribute to one of its Loopback IP addresses. That Loopback IP address is injected into the Service Provider's IGP and, concurrently, a label assigned to it via LDP. Thus, when an ingress LSR is performing a forwarding lookup for a BGP destination it recursively resolves the associated next-hop to a Loopback IP address and associated LDP label of the egress LSR.

Thus, in the context of unlabeled IPv4 traffic, the LDP Entropy Label sub-TLV will typically be applied only to the FEC for the Loopback IP address of the egress LSR and the egress LSR will not announce an entropy label capability for the eBGP learned route.

Authors' Addresses

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

Email: kireeti@juniper.net

John Drake
Juniper Networks
1194 N. Mathilda Ave.
Sunnyvale, CA 94089
US

Email: jdrake@juniper.net

Shane Amante
Level 3 Communications, LLC
1025 Eldorado Blvd
Broomfield, CO 80021
US

Email: shane@level3.net

Wim Henderickx
Alcatel-Lucent
Copernicuslaan 50
2018 Antwerp
Belgium

Email: wim.henderickx@alcatel-lucent.com

Lucy Yong
Huawei USA
1700 Alma Dr. Suite 500
Plano, TX 75075
US

Email: lucyyong@huawei.com

MPLS Working Group
Internet-Draft
Intended status: Standards Track
Expires: September 6, 2011

G. Liu
J. Yang
j. Yu
Z. Fu
ZTE Corporation

March 5, 2011

Multiprotocol Label Switching Transport Profile p2mp Shared Protection
draft-liu-mpls-tp-p2mp-shared-protection-01

Abstract

This document will describle two protection solutions to support protection of failures in p2mp path in MPLS-TP. According to the protection Requirements in RFC 5654, there are requirements for MPLS-TP to support sharing of protection resources such that protection paths that are known not to be required concurrently can share the same resources. In addition, there is a requirement for MPLS-TP to support unidirectional 1:n protection for p2mp paths. These requirements are further addressed in draft-ietf-mpls-tp-survive-fwk . so this draft will present proposed solutions .

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

Status of this Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79. This document may not be modified, and derivative works of it may not be created, and it may not be published except as an Internet-Draft.

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

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

This Internet-Draft will expire on September 6, 2011.

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.

Table of Contents

1. Introduction

This document describes protection solutions for MPLS-TP p2mp paths. The first solution is based on extending 1:1 protection solution to implement 1:n protection by using a shared protection p2mp path when there may have defect in the working p2mp path. A second solution uses a shared p2p bidirectional protection tunnel to protect the branch path of a p2mp working path when detecting defects in a branch path of some p2mp working paths to implement (1:1)^n protection. Both protection solutions satisfy and fulfill requirement 69 and 67B in [RFC 5654]. These solutions can't exclude 1+1 and 1:1 protection solutions for p2mp path in draft-ietf-mpls-tp-survive-fwk and draft-ietf-mpls-tp-linear-protection. it will be used to perfect the requirement of recovery for p2mp path. If only 1+1 protection is used for p2mp path, there need to set up a disjoint protection path for each working path, This will increase the cost of maintaining and monitoring each of these paths (i.e. both the working and protection paths). In addition, since the p2mp service must be transported on both the working and protection paths, more bandwidth resource will be consumed for the p2mp service . Due to these limitations, it is neccesary to consider using shared protection resources for many working paths.

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.

OAM: Operations, Administration, Maintenance

LSP: Label Switched Path.

TLV: Type Length Value

LSR:Label Switching Router

P2MP:Point to Multi-Point

P2P:Point to Point

APS:Automatic Protection Switch

PSC:Protection Switching Coordination

SD:Signal Degrade

SF:Signal Fail

RDI:Remote Defect Indication

MPLS:Multi-Protocol Label Switching

MPLS-TP:Multi-Protocol Label Switching Transport Profile

ME: Maintenance Entity

MEP:MEG End Point

ACH: Associated Channel Header

CC-V: Contunuity Check-Verification;

3. p2mp shared protection solution

This section describes two types of p2mp shared protection solutions. The first proposed solution utilizes one p2mp protection path to protect n p2mp working paths . When a protected p2mp working path detects a defect, the leaf node of the p2mp working path will notify its own root node of defective message by RDI packet through out-of-band return path. If there is no other higher priority protected p2mp working path or control command that requires the use of the protection path, then the defective p2mp service packet will switch to p2mp protection path to be transported. All leaf nodes of the defective p2mp path will select protection path to receive p2mp service packets.

The second proposed solution uses a p2p bidirectional protection tunnel to protect defective branch paths of many protected p2mp working paths. When a defect is detected on a protected branch path of one p2mp working path, the leaf node which has already detected the defect will notify peer node of its own p2p protection path of defective message by extensive APS or PSC packet as the following figure 1. As a result, the service will switch to the protection path to be transported ,so it will bridge both working path and protection path to be transported. But the leaf node of the defective branch path in the p2mp working path will select the protection p2p path to receive the service packet. But other leaf nodes will still receive the service packet from original p2mp working path.

The two p2mp shared protection solutions separately implement 1:n and (1:1)^n protection for p2mp path, The following sub-section describes the protection switching methods in detail.

3.1. 1:n protection

The 1:n protection solution should be similar to 1:1 protection solution described in [survivability-framework] to use one protection path to protect many p2mp service traffics. However, in this mechanism since the protected traffics are transported by different working path. Its implication regarding the p2mp protection path will be configured between the protection domain root node and all leaf nodes of protected p2mp working paths. The operation of this solution is based on the root node of each working p2mp path SHOULD send CC-V OAM packet to all its own leaf nodes periodically. When a leaf node of a p2mp working path fails to receive the CC-V OAM packet for a fixed period, The leaf node should generate RDI packet to notify its own root node of the defective message . When the root node of the p2mp working path receives the RDI packet and knows some failure in one or more one branch path of the p2mp working path, it may send protection switch requirement control packet to the root node of its own protection path and access node of the p2mp service. When the root node of the protection path receives the protection switch requirement control packet from any root node of the protected defective p2mp working path The root node of the protection path MUST choose one defective path to be protected based on the priority of these protected defective p2mp working paths. Then the root node of protection path SHALL generate extensive APS or PSC packet that includes the selected p2mp defective path identifier in a TLV field of the message packet .The following figure 1 is the format of extensive APS or PSC packet .Then it will send the extensive APS or PSC message to all leaf nodes of the p2mp protection path .

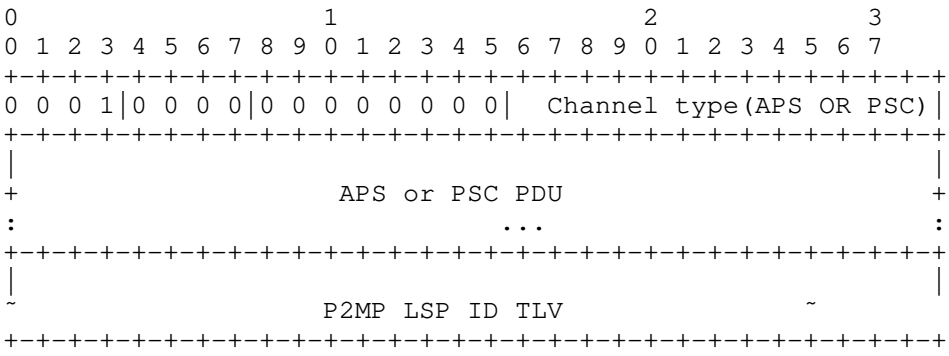


Figure 1

NOTE:

P2MP LSP ID TLV: a standard TLV frame structure. including Type , Length,and Value, and the value field may be identifier of p2mp LSP which have defect and should need to be protected. this p2mp LSP ID TLV format is as the following figure 2

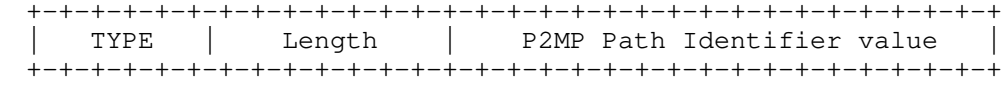
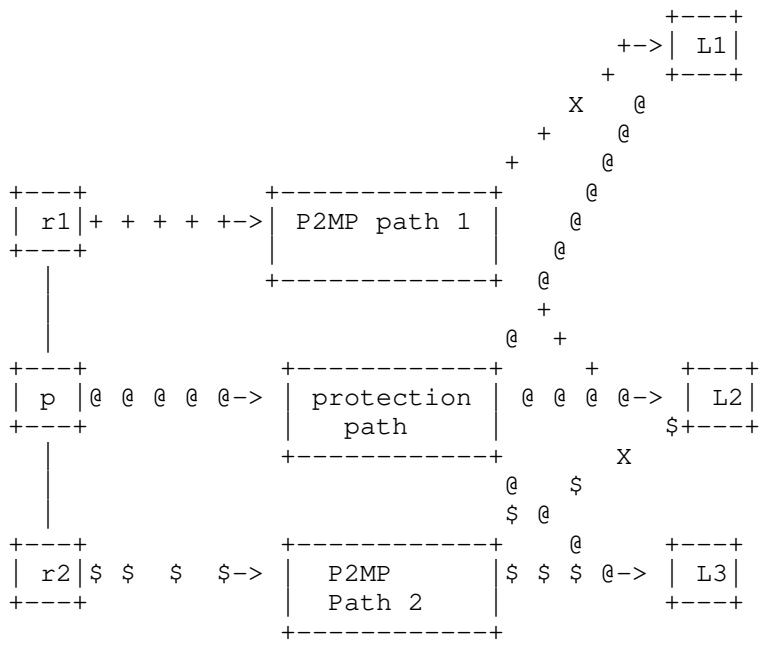


Figure 2

At the same time, the root node of protection path generates notify message control packet and send it to the root node of each defective p2mp working path by control channel ,so the root node of each defective p2mp working path can know whether it may be selected to be protected based on the notify message control packet. If it has already been selected to be protected, it will stop sending service packet in the p2mp working path Then the protected service will switch to the p2mp protection path to be transported.

On the other hand, all leaf nodes of the protection path receive the extension APS or PSC message from the protection path. Then they will know whether to accept and process the service packet from the protection path based on p2mp LSP ID TLV field in the extensive APS or PSC packet. If the leaf node is one sink node of the protected service, it will accept and process the service packet from the protection path. Or else, it will drop the service packet.

the implement in detail as the following figure is 1:2(n=2) protection instance.



NOTE:
@@@@@@: p2mp protection path
+++++: p2mp working path 1
$$$$$: p2mp working path 2
X: failure

Figure 3

For the above p2mp network topology , there are two different p2mp services which need to be transported separately by p2mp working path 1(r1-p2mp path 1-L1,L2) identified by (+) and p2mp working path 2(r2- p2mp path 2-L2,L3)identified by ($) . under normal situation. the p2mp service from root node r1 will be sent and transported to leaf nodes L1,L2 by p2mp working path 1, and another p2mp service from the root node r2 will be sent and transported to leaf nodes L2,L3. in addition, only one p2mp protection path (P-protection path-L1,L2,L3)identified by (@) is used to protect the p2mp working path 1 and the p2mp working path 2. supposing the protection priority of p2mp working path 1 is higher than p2mp working path 2. if there

is a defect in separately branch path(r1-p2mp path 1-L1) of p2mp working path 1 and branch path(r2-p2mp path 2-L2) of p2mp working path 2, Leaf node L1 and leaf node L3 will separately send RDI packet to root node r1 and root node r2 by out-of-band return path. when root node r1 and r2 received the RDI packet and processed it. then the control packet of protection switch requirement will be sent to the root node P of protection path by control channel . Then the root node P will choose one working path to be protected. As the priority of p2mp working path 1 is higher than p2mp working path 2. so the root node P of protection path will select p2mp working path 1 to be protected, and send extensive APS or PSC packet including p2mp LSP ID TLV to all Leaf nodes(L1,L2,L3) of the protection path. At the same time, It will generate response control packet for the protection switch requirement of the root node r1 and r2. the service of the working path 1 will be selected to be protected. So the root node r1 of working path 1 will stop sending its p2mp service in the working path 1, Then the service of the working path 1 will switch to the protection path to be transported. on the other hand, for leaf nodes(L1,L2,L3), when they received the extensive APS or PSC packet from the root node P, They will decide whether to accept and process the service packet from the protection path. As leaf nodes(L1,L2) are the leaf nodes of p2mp working path 1, they will both accept and process the service packet from the p2mp protection path. but for leaf node L3, as it is not the leaf node of p2mp working path 1. it will drop the service packet from the p2mp protection path. While the service of p2mp working path 2 can't be selected to be protected, so the root node r2 will continue to send their own service packet by p2mp working path 2.

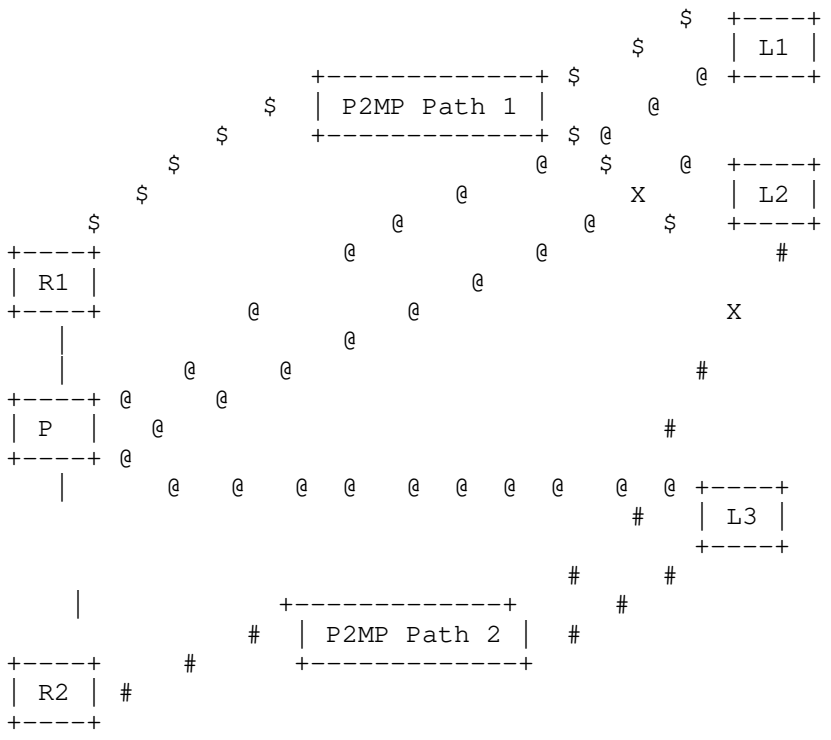
3.2. (1:1)^n protection

This protection solution can use p2p bidirectional protection tunnel to protect some branch paths of many p2mp working paths which have the same leaf node . Under normal situation, the root node of each p2mp working path will periodically send CC-V OAM packet to its own leaf nodes by the p2mp working path to detect defect .In order to protect each branch path of the p2mp working path, Firstly a bidirectional protection p2p path will be pre-configured between source protection node and the leaf node of the protected p2mp working path . in addtion, using an unique adress identifier including IP multicast address,mpls label etc can identify which service is tranported in the protection p2p tunnel. when some leaf nodes detect defect on some branch path of the p2mp working path, it would generate extensive APS or PSC packet Just as the above figure 1 and send it to source protection node . the source protection node received the the packet , it will compare the priority of these defective working path. Then it selects the highest priority service

to be protected and encapsulate the protected service PDU by an unique adress identifier , then it will be sent to the leaf node of the defective branch path.

for example, there is a (1:1)^2(n=2) protection instance as the following figure 4:

there are two p2mp working paths : p2mp working path 1(r1-p2mp Path 1-L1,L2) identified by ($) and p2mp working path 2(r2-p2mp Path 2-L2,L3) identified by (#).in order to protect the service, a bidirectional p2p protection tunnel will be pre-configured between each leaf node(L1,L2,L3) and its own source protection node(P). so it need to configure three p2p protection tunnels identified by (@) in the figure



NOTE:
@: P2P Protection Tunnel
$: P2MP Working Path LSP 1
#: P2MP Working Path LSP 2
X: failure

Figure 4

when the branch path(R1-P2MP Path 1-L2) of p2mp working path LSP 1 and the branch path(R2-P2MP Path 2-L2) of p2mp working path LSP 2 have the defect at the same time, The leaf node L2 will generate extensive APS or PSC packet including p2mp LSP identifier and send it to source protection node(P). while the source protection node(P) received the extensive APS or PSC packet from the Leaf node L2, it will select higher priority service to be protected and encapsulate

the protected service packet by an unique address identifier(IP Multicast address, mpls label etc) and be trasnported through the p2p protection tunnel(P-L2). As the priority of the p2mp working path 1 is higher than the p2mp working path 2. The source protection node (P) will select the service of the p2mp working path 1 to be protected and encapsulate it by the its own unique address identifier . then the protected service of the p2mp working path 1 will be sent by the p2p protection tunnel. at the same time, the leaf node L2 will select the protection tunnel to receive service packet and be based on the unique address identifier to judge which service is transported by p2p protection tunnel now. so that the leaf node L2 can process truely the service packet .

3.3. Conclusion

The two types of p2mp protection solution will individually implement 1:n and (1:1)^n protection for p2mp service. They can fulfill the requirement of unidirectional p2mp protection and sharing protection resource. in addition, the first solution need a special out-of-band return path to send failure message to the root node of p2mp working path. while for the second protection solution, as its protection path is bidirectional , it is unnecessary to set up out-of-band return path for sending failure message. .

4. Security Considerations

The security considerations for the authentication TLV need further study.

5. IANA Considerations

TBD.

6. Acknowledgments

The authors would like to thank yaccov for giving a lot of good comments and revising many syntax for the document. And thank Malcolm Betts and other experts for Providing some good comments and their input to and review of the current document .

7. References

7.1. Normative References

[ITU-T G.8131]
ITU-T, "ITU-T Recommendation(T-MPLS Linear protection)", February 2007.

[RFC 5654]
IETF, "IETF RFC5654(MPLS-TP requirement)", September 2009.

[RFC 5921]
IETF, "IETF RFC5654(MPLS-TP framework)", July 2010.

7.2. Informative References

[L2VPN ICCP]
Luca Martini, Samer Salam, Ali Sajassi, Satoru Matsushima, Matthew Bocci, Thomas D. Nadeau, "Inter-Chassis Communication Protocol for L2VPN PE Redundancy", Oct 2010.

[MPLS-TP Linear protection]
S. Bryant, N. Sprecher, H. van Helvoort,A. Fulignoli Y. Weingarten, "MPLS transport profile Linear Protection", July 2010.

[MPLS-TP Survivability Framework]
N. Sprecher, A. Farrel, "Multiprotocol Label Switching Transport Profile Survivability Framework", June 2010.

[MPLS-TP linear protection]
Z.Haiyan, I.umansky, L. han,J.Ryoo, D'Alessandro , "Linear Protection Switching in MPLS-TP", July 2010.

7.3. URL References

[MPLS-TP-22]
IETF - ITU-T Joint Working Team, "", 2008, <http://www.example.com/dominator.html>.

Authors' Addresses

Liu guoman
ZTE Corporation
No.68, Zijinghua Road, Yuhuatai District
Nanjing 210012
P.R.China

Phone: +86 025 52871606
Email: liu.guoman@zte.com.cn

Yang Jian
ZTE Corporation
5F,RD Building 3,ZTE Industrial Park,XiLi LiuXian Road
Nanshan District,Shenzhen 518055
P.R.China

Phone: +86 755 26773731
Email: yang_jian@zte.com.cn

Yu jinghai
ZTE Corporation
No.68, Zijinghua Road, Yuhuatai District
Nanjing 210012
P.R.China

Phone: +86 025 52871606
Email: yu.jinghai@zte.com.cn

Fu zhentao
ZTE Corporation
No.68, Zijinghua Road, Yuhuatai District
Nanjing 210012
P.R.China

Phone: +86 025 52871745
Email: fu.zhentao@zte.com.cn

Internet Engineering Task Force
Internet Draft
Intended status: Standards Track
Expires: September 2011

Luca Martini
George Swallow
Cisco

March 13, 2011

MPLS LSP PW status refresh reduction for Static Pseudowires

draft-martini-pwe3-status-aggregation-protocol-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/1id-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 September 13, 2010

Abstract

This document describes a method includes generating an aggregated pseudowire status message on Multi-Protocol Label Switching (MPLS) network Label Switched Path (LSP). The method for transmitting the pseudowire (PW) status information is not new, however these protocol extension allows a Service Provider (SP) to reliably use the PW static status messages on individual PWs. The aggregated pseudowire status message configured to verify a current status of all pseudowires on the LSP.

Table of Contents

1. Introduction

When PWs use an Multi Protocol Label Switched (MPLS) network as the Packet Switched Network (PSN), are setup according to [RFC4447] static configuration mode, the PW status information is propagated using the method described in [PW-STATUS]. There are 2 basic modes of operation described in [PW-STATUS] section 5.3: Periodic retransmission of non-zero status messages, and a simple acknowledge of PW status (sec 5.3.1 of [PW-STATUS]). The LSP level protocol described below applies to the case then PW status is acknowledged

immediately with a requested refresh value of zero. (no refresh) In this case the PW status refresh reduction protocol is necessary for several reasons , such as:

-i. Greatly increase the scalability of the PW status protocol by reducing the amount of messages that a PE needs to periodically send to it's neighbors.
-ii. Detect a remote PE restart.
-iii. If the local state is lost for some reason, the PE needs to be able to request a status refresh from the remote PE
-iv. Optionally detect a remote PE provisioning change.

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 [RFC2119].

1.2. Terminology

FEC: Forwarding Equivalence Class

LDP: Label Distribution Protocol

LSP: Label Switching Path

MS-PW: Multi-Segment Pseudowire

PE: Provider Edge

PW: Pseudowire

SS-PW: Single-Segment Pseudowire

S-PE: Switching Provider Edge Node of MS-PW

T-PE: Terminating Provider Edge Node of MS-PW

1.3. Notational Conventions in Backus-Naur Form

All multiple-word atomic identifiers use underscores (_) between the words to join the words. Many of the identifiers are composed of a concatenation of other identifiers. These are expressed using Backus-Naur Form (using double-colon - "::" - notation).

Where the same identifier type is used multiple times in a concatenation, they are qualified by a prefix joined to the identifier by a dash (-). For example Src-Node_ID is the Node_ID of a node referred to as Src (where "Src" is short for "source" in this example).

The notation does not define an implicit ordering of the information elements involved in a concatenated identifier.

2. PW status refresh reduction protocol

PW status refresh reduction protocol consists of a simple message that is sent at the LSP level using the MPLS Generic Associated Channel.

A PE using the PW status refresh reduction protocol MUST send the PW status refresh reduction Message as soon as a PW is configured on a particular LSP. The message is then re-transmitted at a locally configured interval indicated in the refresh timer field. If no acknowledgment is received, the protocol does not reach active state, and the PE SHOULD NOT send any PW status messages with a refresh timer of zero as described in [PW-STATUS] section 5.3.1.

2.1. Protocol states

The protocol can be in 3 possible states: INACTIVE, STARTUP, and ACTIVE.

2.1.1. INACTIVE

This state is entered when the protocol is turned off. This state is also entered if all PW on a specific LSP are unprovisioned, or the feature is unprovisioned.

2.1.2. STARTUP

In this state the PE transmits periodic PW status refresh messages, with the Ack Session ID set to 0. The PE remains in this state until a PW status refresh message is received with the correct local session ID in the Ack Session ID Field. This state can be exited to the ACTIVE or INACTIVE state.

2.1.3. ACTIVE

This state is entered once the PE receives a PW status refresh message with the correct local session ID in the Ack Session ID Field within 3.5 times the refresh timer field value of the last PW status refresh message transmitted. This state is immediately exited as follows:

- -i. A valid PW status refresh message is not received within 3.5 times the current refresh timer field value. (assuming a timer transition procedure is not in progress) New state: STARTUP
- -ii. A PW status refresh message is received with the wrong, or a zero, Ack Session ID field value. New state: STARTUP
- -iii. All PWs using the particular LSP are unprovisioned, or the protocol is disabled. New state: INACTIVE

2.2. Timer value change transition procedure

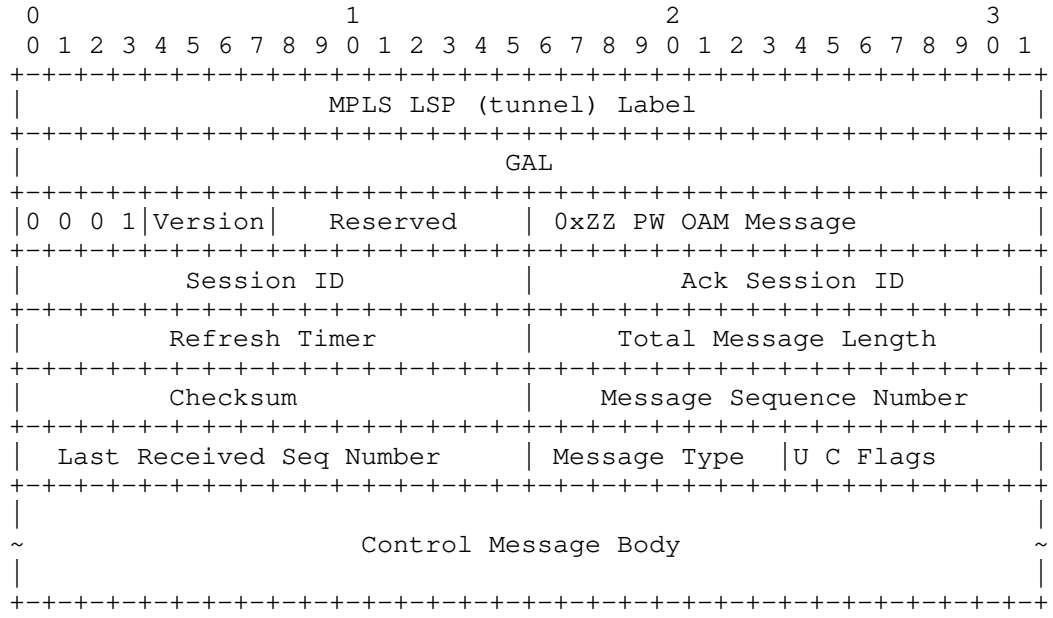
If a PE needs to change the refresh timer value field while the PW refresh reduction protocol is in the ACTIVE state, the following procedure must be followed:

- -i. A PW status refresh message is transmitted with the new timer value.
- -ii. If the new value is greater then the original one the PE will operate on the new timer value immediately.
- -iii. If the new value is smaller then the original one, the PE will operate according to the original timer value for a period 3.5 times the original timer value, or until the first valid PW status refresh message is received.

  A PE receiving a PW status refresh message with a new timer value, will immediately transmit an acknowledge PW status refresh message, and start operating according to the new timer value.

3. PW status refresh reduction Message Encoding

The packet containing the refresh reduction message is encoded as follows: (omitting link layer information)



This message contains the following fields:

* PW OAM Message.

  This field indicates the generic associated channel type in the GACH header as defined in [RFC5586].

  Note: Channel type 0xZZ pending IANA allocation.

* Session ID

  A non-zero, locally selected session number that is not preserved if the local PE restarts.

* Ack Session ID

  The Acknowledgment Session ID received from the remote PE.

* Refresh Timer.

  A non zero unsigned 16 bit integer value greater or equal to 10, in milliseconds, that indicates the desired refresh interval. The default value of 30000 is RECOMENDED.

* Total Message Length

  Total length in octets of the Checksum, Message Type, Flags, Message Sequence Number, and control message body. A value of zero means no control message is present, and therefore no Checksum, and following fields are present either.

* Checksum

  A 16 bit field containing the one's complement of the one's complement sum of the entire message (including the GACH header), with the checksum field replaced by zero for the purpose of computing the checksum. An all-zero value means that no checksum was transmitted. Note that when the checksum is not computed, the header of the bundle message will not be covered by any checksum.

* Message Sequence Number

  A unsigned 16 bit integer number that is started from 1 when the protocol enters ACTIVE state. The sequence numbers wraps back to 1 when the maximum value is reached. The value of zero is reserved and MUST NOT be used.

* Last Received Message Sequence Number

  The sequence number of the last message received. In no message has yet been received during this session, this field is set to zero.

* Message Type

  The Type of the control message that follows. Control message types are allocated in this document, and by IANA.

* (U) Unknown flag bit.

  Upon receipt of an unknown message, if U is clear (=0), the keepalive session MUST be terminated by entering STARTUP state; if U is set (=1), the unknown message MUST be acknowledge and silently ignored and the following messages, if any, processed as if the unknown message did not exist.

* (C) Configuration flag bit. The C Bit is used to signal the end of PW configuration transmission. If it is set, the sending PE has finished sending all it's current configuration information.

* Flags (Reserved)

  7 bits of flags reserved for future use, they MUST be set to 0 on transmission, and ignored on reception.

* Control Message Body

  The Control Message body is defined in a section below, and is specific to the type of message.

It should be noted that the Checksum, Message Sequence Number, Last Received Message Sequence Number, Message Type, Flags, and control message body are OPTIONAL.

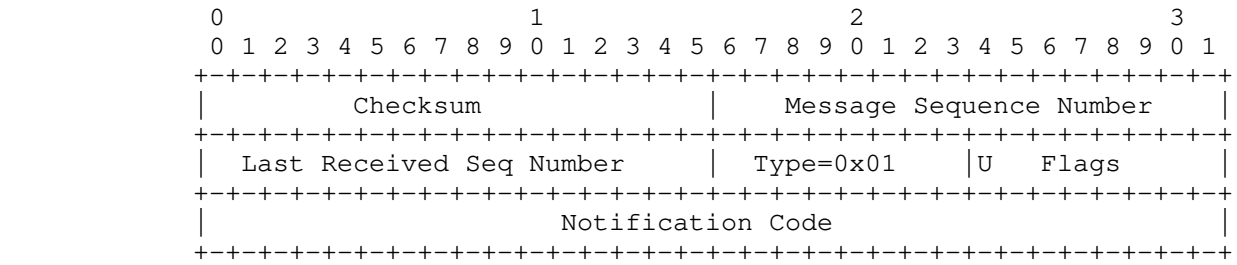
4. PW status refresh reduction Control Messages

PW status refresh reduction Control messages consist of the Checksum, Message Sequence Number, Last Received Message Sequence Number, Message Type, Flags, and control message body. There can only be one control message construct per PW status refresh reduction Message. If the U bit is set, and a PE receiving the PW status refresh reduction Message does not understand the control message, the control message MUST be silently ignored. However the control message sequence number MUST still be acknowledged by sending a null message back with the appropriate value in the Last Message Received Field. If a control message is not acknowledge, after 3.5 times the value of the Refresh Timer, a fatal notification "unacknowledged control message" MUST be sent, and the PW refresh reduction session MUST be terminated.

If a PE does not want or need to send a control message, the Checksum , and all following fields MUST NOT be sent, and the Total Message Length field is then set to zero.

4.0.1. Notification message

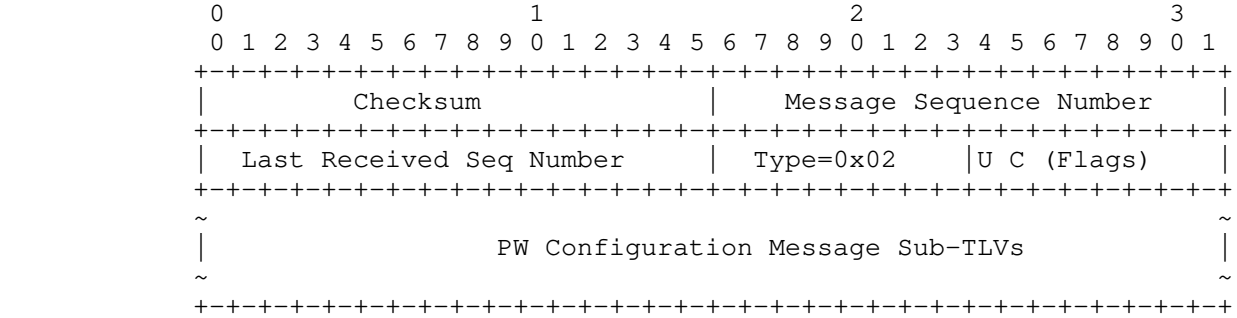
The most common use of the Notification Message is to acknowledge the reception of a message by indicating the received message sequence number in the "Last Received Sequence Number" field. The notification message is encoded as follows:



The message type is set to 0x01, and the U bit is treated as described in the above section. The Notification Codes are a 32 bit quantity assigned by IANA. (see IANA consideration section) Notification codes are either are either considered "Error codes" or simple notifications. If the Notification code is an Error code as indicated in the IANA allocation registry, the keepalive session MUST be terminated by entering STARTUP state.

4.0.2. PW Configuration Message

The PW status refresh reduction TLVs are informational TLVs, that allow the remote PE to verify certain provisioning information. This message contain a series of sub-TLVs in no particular order, that contain PW ,and LSP configuration information. The message has no preset length limit, however its total length will be limited by the transport network Maximum Transmit Unit (MTU).



The PW Configuration Message type is set to 0x02. For this message the U-bit is set to 1 as processing of these messages is OPTIONAL.

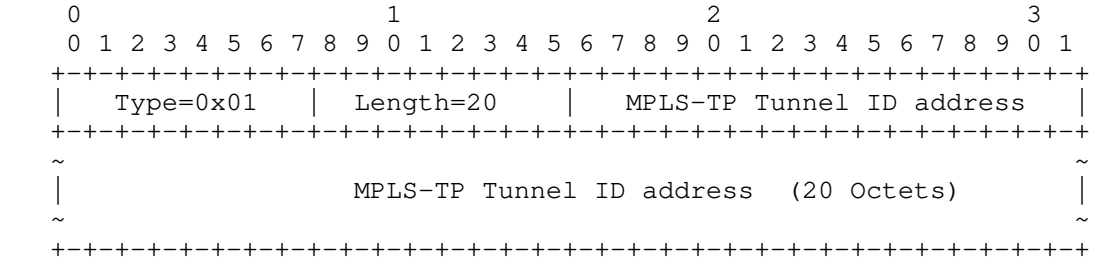
The C Bit is used to signal the end of PW configuration transmission. If it is set, the sending PE has finished sending all it’s current configuration information. The PE transmitting the configuration MUST

set the C bit on the last PW configuration message when all current PW configuration has been sent.

4.0.2.1. MPLS-TP Tunnel ID

This TLV contains the address of the MPLS-TP tunnel ID. When the configuration message is used for a particular keepalive session the MPLS-TP Tunnel ID sub-TLV MUST be sent at least once.

The MPLS-TP Tunnel ID address is encoded as follows:



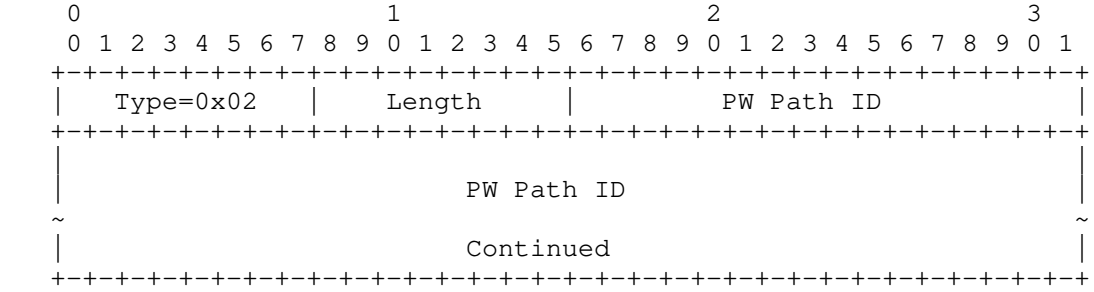
The MPLS-TP point to point tunnel ID is defined in [IDENTIFIER] as follows:

Src-Global_Node_ID::Src-Tunnel_Num::Dst-Global_Node_ID::Dst-Tunnel_Num

Note that a single address is enough to identify the tunnel, and the source end of the message.

4.0.2.2. PW ID configured List

This OPTIONAL TLV contains a list of the provisioned PWs on the LSP.

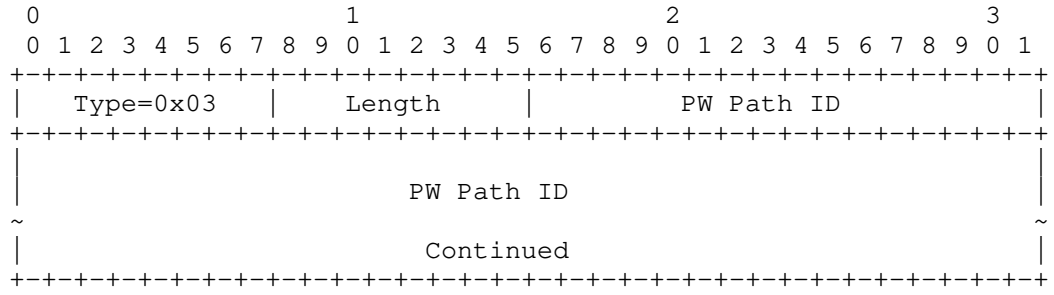


The PW Path ID is a 32 octet pseudowire path identifier specified in [IDENTIFIER] as follows: AGI::Src-Global_ID::Src-Node_ID::Src-AC_ID:: Dst-Global_ID::Dst-Node_ID::Dst-AC_ID

The number of PW Path IDs in the TLV will be inferred by the length of the TLV up to a maximum of 8. The procedure for processing this TLV will be described in a section below.

4.0.2.3. PW ID unconfigured List

This OPTIONAL TLV contains a list of the PWs that have been unprovisioned on the LSP. Note that it is a fatal session error to send the same PW address in both the configured list TLV , and the unconfigured list TLV in the same configuration message.



The PW Path ID is a 32 octet pseudowire path identifier specified in [IDENTIFIER] as follows: AGI::Src-Global_ID::Src-Node_ID::Src-AC_ID:: Dst-Global_ID::Dst-Node_ID::Dst-AC_ID

The number of PW Path IDs in the TLV will be inferred by the length of the TLV up to a maximum of 8.

5. PW provisioning verification procedure

This procedure , and the advertisement of the PW configuration message are OPTIONAL.

A PE that desires to use the PW configuration message to verify the configuration of PWs on a particular LSP, should advertise it’s PW configuration to the remote PE on LSPs that have active keepalive sessions. When a PE receives PW configuration information using this protocol, and it not supporting, or not willing to use the information it MUST acknowledge all the PW configuration message, with a notification of "PW configuration not supported.". In this

case, the information in the control messages is silently ignored. If a PE receives such a notification it should stop sending PW configuration control messages for the duration of the PW refresh reduction keepalive session.

If PW configuration information is received, it is used to verify the accuracy of the local configuration information against the remote PE's configuration information. If a configuration mismatch is detected, where a particular PW is configured locally, but not on the remote PE the following action SHOULD be taken:

-i. The local PW MUST be considered in "Not Forwarding" State.

-ii. The PW Attachment Circuit status is set to reflect the PW fault.

-iii. An Alarm MAY be raised to a network management system.

5.1. PW ID List advertising and processing

When configuration messages are advertised along a particular LSP, the PE sending the messages needs to check point the configuration information sent by setting the C bit when all currently known configuration information has been sent. This process allows the receiving PE to immediately proceed to verify all the currently configured PWs on that LSP, eliminating the need for a long waiting period.

If a new PW is added to a particular LSP, the PE MUST place the configuration verification of this PW on hold for a period of at least 10 seconds. This is necessary to prevent false positive events of mis-configuration due to the ends of the PW being slightly out of sync.

6. PW status refresh procedure

When the the refresh reduction protocol, on a particular LSP, is in the ACTIVE state, the PE can send all PW status messages, for PWs on that LSP, with a refresh timer value of zero. This greatly decreases the amount of messages that the PE needs to transmit to the remote PE because once the PW status message for a particular PW is acknowledged, further repetitions of that message are no longer necessary.

To further mitigate the amount of possible messages when an LSP starts forwarding traffic, care should be taken to permit the PW

refresh reduction protocol to reach the ACTIVE state quickly, and before the the first PW status refresh timer expires. This can be achieved by using a PW status refresh reduction Message refresh timer value that is much smaller then the PW status message refresh timer value in use. (sec 5.3.1 of [PW-STATUS])

If the refresh reduction protocol session is terminated by entering the INACTIVE or STARTUP states, the PE MUST immediately re-send all the previously sent PW status messages for that particular LSP for which the session terminated. In this case the refresh timer value MUST NOT be set to zero, and MUST be set according to the local policy of the PE router.

7. Security Considerations

Section to be completed in a later version of the document.

8. IANA Considerations

8.1. PW Status Refresh Reduction Message Types

IANA needs to set up a registry of "PW status refresh reduction Control Messages". These are 8-bit values. Type value 1 through 2 are defined in this document. Type values 3 through 64 are to be assigned by IANA using the "Expert Review" policy defined in RFC5226. Type values 65 through 127, 0 and 255 are to be allocated using the IETF consensus policy defined in [RFC5226]. Type values 128 through 254 are reserved for vendor proprietary extensions and are to be assigned by IANA, using the "First Come First Served" policy defined in RFC5226.

The Type Values are assigned as follows:

Type	Message Description
0x01	Notification message
0x02	PW Configuration Message

8.2. PW Configuration Message Sub-TLVs

IANA needs to set up a registry of "PW status refresh reduction Configuration Message Sub-TLVs". These are 8-bit values. Type value 1 through 2 are defined in this document. Type values 3 through 64 are to be assigned by IANA using the "Expert Review" policy defined in RFC5226. Type values 65 through 127, 0 and 255 are to be allocated using the IETF consensus policy defined in [RFC5226]. Type values 128

through 254 are reserved for vendor proprietary extensions and are to be assigned by IANA, using the "First Come First Served" policy defined in RFC5226.

The Type Values are assigned as follows:

sub-TLV type	Description
0x01	MPLS-TP Tunnel ID address.
0x02	PW ID configured List.
0x03	PW ID unconfigured List.

8.3. PW Status Refresh Reduction Notification Codes

IANA needs to set up a registry of "PW status refresh reduction Notification Codes". These are 32-bit values. Type value 1 through 7 are defined in this document. Type values 8 through 65536 are to be assigned by IANA using the "Expert Review" policy defined in RFC5226. Type values 65536 through 134,217,728, 0 and 4,294,967,295 are to be allocated using the IETF consensus policy defined in [RFC5226]. Type values 134,217,729 through 4,294,967,294 are reserved for vendor proprietary extensions and are to be assigned by IANA, using the "First Come First Served" policy defined in RFC5226.

The Type Values are assigned as follows: nf Code Error? Description ---- ------ ----------- 0x00000000 No Null Notification. 0x00000001 No PW configuration rejected. 0x00000002 Yes PW Configuration TLV conflict. 0x00000003 No Unknown TLV (U-bit=1) 0x00000004 Yes Unknown TLV (U-bit=0) 0x00000005 No Unknown Message Type 0x00000006 No PW configuration not supported. 0x00000007 Yes Unacknowledged control message.

9. References

9.1. Normative References

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

[RFC4447] "Transport of Layer 2 Frames Over MPLS", Martini, L., et al., rfc4447 April 2006.

[PW-STATUS] L. Martini, G. Swallow, G. Heron, M. Bocci "Pseudowire Status for Static Pseudowires", draft-ietf-pwe3-static-pw-status-03.txt, (work in progress),

March 2011

[IDENTIFIER] M. Bocci, G. Swallow, E. Gray "MPLS-TP Identifiers" draft-ietf-mpls-tp-identifiers-04.txt, IETF Work in Progress, March 2011

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

9.2. Informative References

[RFC5586] M. Bocci, Ed., M. Vigoureux, Ed., S. Bryant, Ed., "MPLS Generic Associated Channel", rfc5586, June 2009

10. Author's Addresses

Luca Martini
Cisco Systems, Inc.
9155 East Nichols Avenue, Suite 400
Englewood, CO, 80112
e-mail: lmartini@cisco.com

George Swallow
Cisco Systems, Inc.
300 Beaver Brook Road
Boxborough, Massachusetts 01719
United States
e-mail: swallow@cisco.com

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.

Expiration Date: September 2011

IETF
Internet Draft

Ping Pan
Mohana Srinivas
Rajan Rao
Biao Lu
(Infinera)
Sam Aldrin
(Huawei)

Expires: September 14, 2011 March 14, 2011

Supporting Shared Mesh Protection in MPLS-TP Networks

draft-pan-shared-mesh-protection-01.txt

Status of this Memo

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

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79. This document may not be modified, and derivative works of it may not be created, and it may not be published except as an Internet-Draft.

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79. This document may not be modified, and derivative works of it may not be created, except to publish it as an RFC and to translate it into languages other than English.

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.

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/1id-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 September 14, 2011.

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.

Abstract

Shared mesh protection is a common protection and recovery mechanism in transport networks, where multiple paths can share the same set of network resources for protection purposes.

In the context of MPLS-TP, it has been explicitly requested as a part of the overall solution (Req. 67, 68 and 69 in RFC5654 [1]).

It's important to note that each MPLS-TP LSP may be associated with transport network resources. In event of network failure, it may require explicit activation on the protecting paths before switching user traffic over.

In this memo, we define a lightweight signaling mechanism for protecting path activation in shared mesh protection-enabled MPLS-TP networks.

Table of Contents

1. Introduction

Shared mesh protection is a common protection and recovery mechanism in transport networks, where multiple paths can share the same set of network resources for protection purposes.

In the context of MPLS-TP, it has been explicitly requested as a part of the overall solution (Req. 67, 68 and 69 in RFC5654 [1]).Its operation has been further outlined in Section 4.7.6 of MPLS-TP Survivability Framework [2].
It's important to note that each MPLS-TP LSP may be associated with transport network resources. In event of network failure, it may require explicit activation on the protecting paths before switching user traffic over.

In this memo, we define a lightweight signaling mechanism for protecting path activation in shared mesh protection-enabled MPLS-TP networks.

Here are the key design goals:

1. Fast: The protocol is to activate the previously configured protecting paths in a timely fashion, with minimal transport and processing overhead. The goal is to support 50msec end-to-end traffic switch-over in large transport networks.

2. Reliable message delivery: Activation and deactivation operation have serious impact on user traffic. This requires the protocol to adapt a low-overhead reliable messaging mechanism.

3. Modular: Depending on deployment scenarios, the signaling may need to support functions such as preemption, resource re-allocation and bi-directional activation in a modular fashion.

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

2. Background

Transport network protection can be typically categorized into three types:

Cold Standby: In this type of protection, the nodes will only negotiate and establish backup path after the detection of network failure.

Hot Standby: The protecting paths are established prior to network failure. This is also known as "make-before-break". Upon the detection of network failure, the edge nodes will switch data traffic into pre-established backup path immediately.

Warm Standby: The nodes will negotiate and reserve protecting path prior to network failure. However, data forwarding path will not be programmed. Upon the detection of network failure, the nodes will send explicit messages to relevant nodes to "wake up" the protecting path.

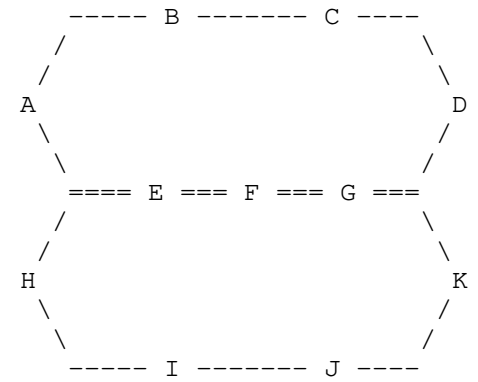
The activation signaling defined in this memo is to support warm standby in the context of MPLS-TP.

Further, the activation procedure may be triggered using the failure notification methods defined in MPLS-TP OAM specifications.

3. Problem Definition

In this section, we describe the operation of shared mesh protection in the context of MPLS-TP networks, and outline some of the relevant definitions.

We refer to the figure below for illustration:



Working paths: X = {A, B, C, D}, Y = {H, I, J, K}

Protecting paths: X' = {A, E, F, G, D}, Y' = {H, E, F, G, K}

The links between E, F and G are shared by both protecting paths. All paths are established via MPLS-TP control plane prior to network failure.

All paths are assumed to be bi-directional. An edge node is denoted as a headend or tailend for a particular path in accordance to the path setup direction.

Initially, the operators setup both working and protecting paths. During setup, the operators specify the network resources for each path.

The working path X and Y will configure the appropriate resources on the intermediate nodes, however, the protecting paths, X' and Y', will reserve the resources on the nodes, but won't occupy them.

Depending on network planning requirements (such as SRLG), X' and Y' may share the same set of resources on node E, F and G. The resource

assignment is a part of the control-plane CAC operation taking place on each node.

At some time, link B-C is cut. Node A will detect the outage, and initiate activation messages to bring up the protecting path X'. The intermediate nodes, E, F and G will program the switch fabric and configure the appropriate resources. Upon the completion of the activation, A will switch the user traffic to X'.

The operation may have extra caveat:

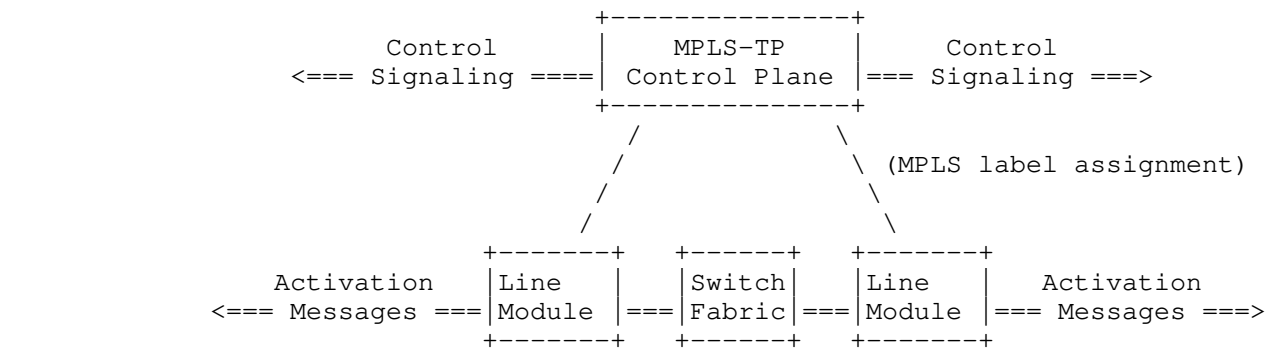
1. Preemption: Protecting paths X' and Y' may share the same resources on node E, F or G due to resource constraints. Y' has higher priority than that of X'. In the previous example, X' is up and running. When there is a link outage on I-J, H can activate its protecting path Y'. On E, F or G, Y' can take over the resources from X' for its own traffic. The behavior is acceptable with the condition that A should be notified about the preemption action.

2. Over-subscription (1:N): A unit of network resource may be reserved by one or multiple protecting paths. In the example, the network resources on E-F and F-G are shared by two protecting paths, X' and Y'. In deployment, the over-subscription ratio is an important factor on network resource utilization.

4. Protection Switching

The entire activation and switch-over operation need to be within the range of milliseconds to meet customer's expectation [1]. This section illustrates how this may be achieved on MPLS-TP-enabled transport switches. Note that this is for illustration of protection switching operation, not mandating the implementation itself.

The diagram below illustrates the operation.



Typical MPLS-TP user flows (or, LSP's) are bi-directional, and setup as co-routed or associated tunnels, with a MPLS label for each of the upstream and downstream traffic. On this particular type of transport switch, the control-plane can download the labels to the line modules. Subsequently, the line module will maintain a label lookup table on all working and protecting paths.

Upon the detection of network failure, the headend nodes will transmit activation messages along the MPLS LSP's. When receiving the messages, the line modules can locate the associated protecting path from the label lookup table, and perform activation procedure by programming the switching fabric directly. Upon its success, the line module will swap the label, and forward the activation messages to the next hop.

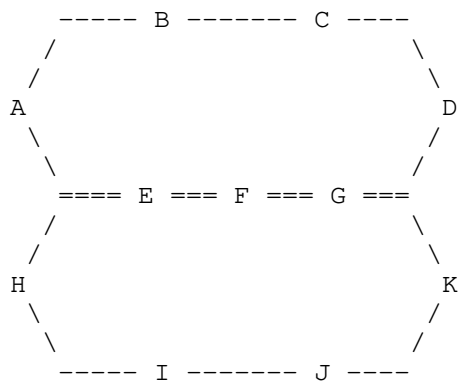
In summary, the activation procedure involves efficient path lookup and switch fabric re-programming.

To achieve the tight end-to-end switch-over budget, it's possible to implement the entire activation procedure with hardware-assistance (such as in FPGA or ASIC).

The activation messages are encapsulated with a MPLS-TP Generic Associated Channel Header (GACH) [3]. Detailed message encoding is explained in Section 6.

5. Activation Operation Overview

In this section, we describe the activation procedure using the same figure shown before:



Working paths: X = {A, B, C, D}, Y = {H, I, J, K}

Protecting paths: X' = {A, E, F, G, D}, Y' = {H, E, F, G, K}

Upon the detection of working path failure, the edge nodes, A, D, H and K may trigger the activation messages to activate the protecting paths, and redirect user traffic immediately after.

We assume that there is a consistent definition of priority levels among the paths throughout the network. At activation time, each node may rely on the priority levels to potentially preempt other paths.

When the nodes detect path preemption on a particular node, they should inform all relevant nodes to free the resources.

To optimize traffic protection and resource management, each headend should periodically poll the protecting paths about resource availability. The intermediate nodes have the option to inform the current resource utilization.

Note that, upon the detection of a working path failure, both headend and tailend may initiate the activation simultaneously (known as bi-directional activation). This may expedite the activation time. However, both headend and tailend nodes need to coordinate the order of protecting paths for activation, since there may be multiple protecting paths for each working path (i.e., 1:N protection). For clarity, we will describe the operation from headend in the memo. The tailend operation will be available in the subsequent revisions.

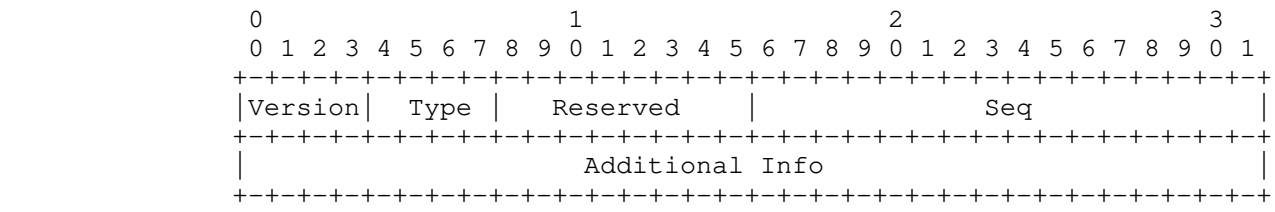
6. Protocol Definition

6.1. Activation Messages

The activation requires the following messages:

- ENABLE: this is initiated by the headend nodes to activate a protecting path
- DISABLE: this is initiated by the headend nodes to disable a protecting path and free the associated network resources
- GET: this is initiated by the headend to gather resource availability information on a particular protecting path
- NOTIFY: this is initiated by the intermediate nodes and terminate on the headend nodes to report preemption or protection failure conditions
- STATUS: this is the acknowledgement message for ENABLE, DISABLE, GET, and NOTIFY messages, and contains the relevant status information

Each activation message has the following format:



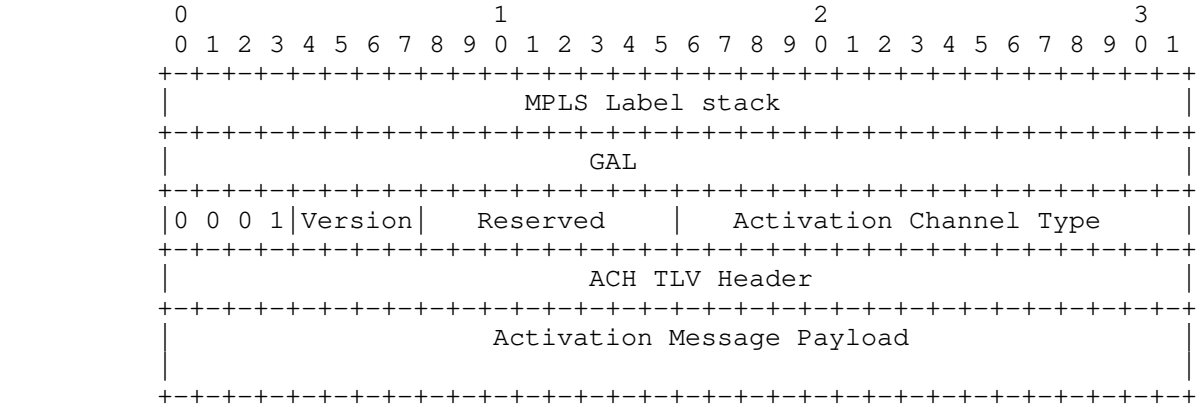
- Version: 1
- Type:
  - ENABLE 1
  - DISABLE 2
  - GET 3
  - STATUS 4

o NOTIFY 5

- Reserved: This field is reserved for future use
- Seq: This uniquely identifies a particular message. This field is defined to support reliable message delivery
- Additional Info: the message-specific data

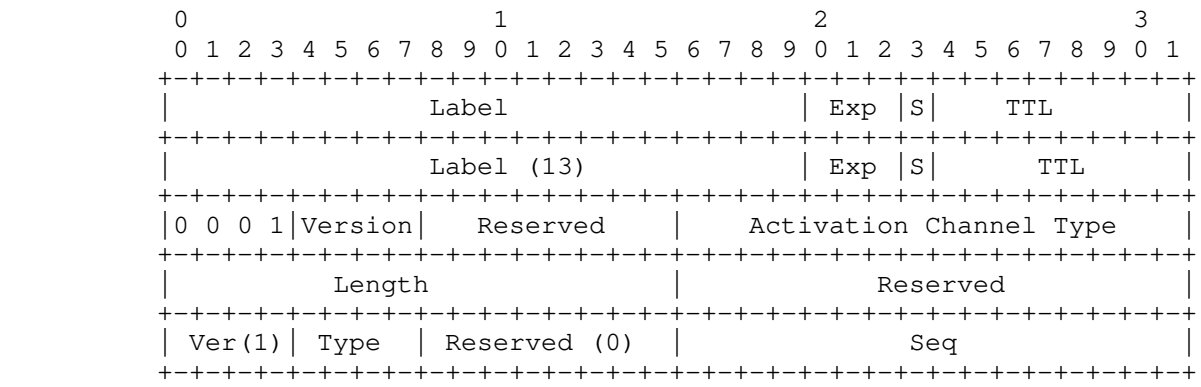
6.2. Message Encapsulation

Activation messages use MPLS labels to identify the paths. Further, the messages are encapsulated in GAL/GACH:

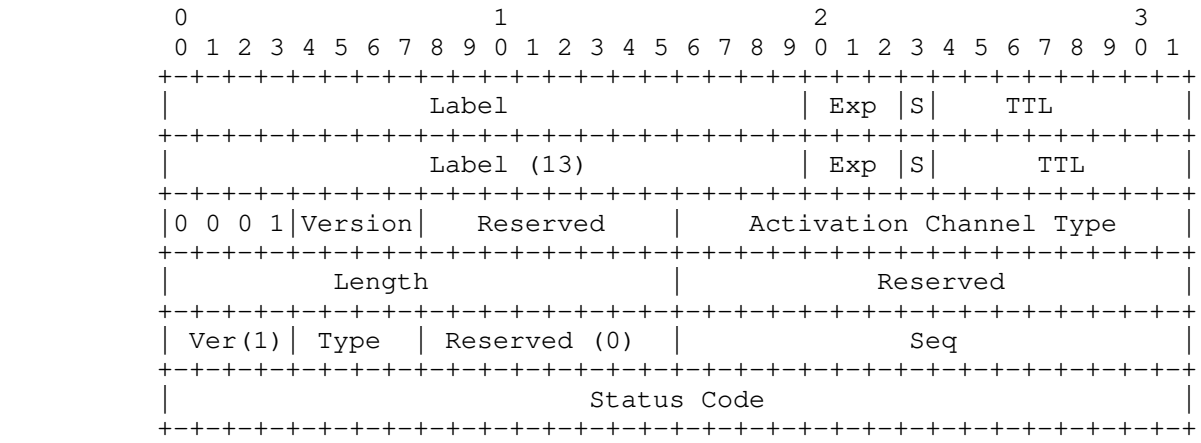


- GAL is described in [3]
- Activation Channel Type is the GACH channel number assigned to the protocol. This uniquely identifies the activation messages.
- ACH TLV Header contains the message length, and is described in [3]

Specifically, ENABLE, DISABLE and GET messages have the following message format:



Both STATUS and NOTIFY messages have the following message format:



Currently, the status code used for acknowledging and preemption notification has the following definition:

- o 1xx: OK
  - . 101: end-to-end ack
- o 2xx: message processing errors
  - . 201: no such path
- o 3xx: processing issues:

. 301: no more resource for the path

. 302: preempted by another path

. 303: system failure

o 4xx: informative data:

. 401: shared resource has been taken by other paths

Further, for preemption notification, we may consider of using the existing MPLS-TP OAM messaging. More details will be available in the future revisions.

6.3. Reliable Messaging

The activation procedure adapts a simple two-way handshake reliable messaging.

Each node maintains a sequence number generator. Each new sending message will have a new sequence number. After sending a message, the node will wait for a response with the same sequence number.

Specifically, upon the generation of ENABLE, DISABLE, GET and NOTIFY messages, the message sender expects to receive a STATUS in reply with same sequence number.

If a sender is not getting the reply (STATUS) within a time interval, it will retransmit the same message with a new sequence number, and starts to wait again. After multiple retries (by default, 3), the sender will declare activation failure, and alarm the operators for further service.

6.4. Message Scoping

Activation signaling uses MPLS label TTL to control how far the message would traverse. Here are the processing rules on each intermediate node:

- On receive, if the message has label TTL = 0, the node must drop the packet without further processing
- The receiving node must always decrement the label TTL value by one. If TTL = 0 after the decrement, the node must process the message. Otherwise, the node must forward the message without further processing (unless, of course, the node is headend or tailend)

- On transmission, the node will adjust the TTL value. For hop-by-hop messages, TTL = 1. Otherwise, TTL = 0xFF, by default.

7. Processing Rules

7.1. Enable a protecting path

Upon the detection of network failure on a working path, the headend node identifies the corresponding MPLS-TP label and initiates the protection switching by sending an ENABLE message.

ENABLE messages always use MPLS label TTL = 1 to force hop-by-hop process. Upon reception, a next-hop node will locate the corresponding path and activate the path.

If the Enable message is received on an intermediate node, due to label TTL expiry, the message is processed and then propagated to the next hop of the MPLS TP LSP, by setting the MPLS TP label TTL = 1. The intermediate node may NOT respond back to the headend node with STATUS message.

The headend node will declare the success of the activation only when it gets a positive reply from the tailend node. This requires that the tailend nodes must reply STATUS messages to the headend nodes in all cases.

If the headend node is not receiving the acknowledgement within a time internal, it will retransmit another ENABLE message with a different Seq number.

If the headend node is not receiving a positive reply within a longer time interval, it will declare activation failure.

If an intermediate node cannot activate a protecting path, it will reply an NOTIFY message to report failure. When the headend node receives a NOTIFY message for failure, it must initiate DISABLE messages to clean up networks resources on all the relevant nodes on the path.

7.2. Disable a protecting path

The headend removes the network resources on a path by sending DISABLE messages.

In the message, the MPLS label represents the path to be de-activated. The MPLS TTL is one to force hop-by-hop processing.

Upon reception, a node will de-activate the path, by freeing the resources from the data-plane.

As a part of the clean-up procedure, each DISABLE message must traverse through and be processed on all the nodes of the corresponding path. When the DISABLE message reaches to the tailend node, the tailend is required to reply with a STATUS message to the headend.

The de-activation process is complete when the headend receives the corresponding STATUS message from the tailend.

7.3. Get protecting path status

The operators have the option to trigger GET messages from the headend to check on the protecting path periodically or on-demand. The process procedure on each node is very similar to that of ENABLE messages on the intermediate nodes, except the GET messages should not trigger any network resource re-programming.

Upon reception, the node will check the availability of resources.

If the resource is no longer available, the node will reply a NOTIFY with error conditions.

7.4. Acknowledgement with STATUS

The STATUS message is the acknowledgement packet to all messages, and may be generated by any node in the network.

Each STATUS message must use the same sequence number as the corresponding message (ENABLE, DISABLE, GET and NOTIFY).

When replying to headend, the tailend nodes must originate STATUS messages with a large MPLS TTL value (0xff, by default).

7.5. Preemption

The preemption operation typically takes place when processing an ENABLE message.

If the activating network resources have been used by another path and carrying user traffic, the node needs to compare the priority levels.

If the existing path has higher priority, the node needs to reject the ENABLE message by sending a STATUS message to the corresponding headend to inform the unavailability of network resources.

If the new path has higher priority, the node will reallocate the resource to the new path, and send an NOTIFY message to old path's headend node to inform about the preemption.

8. Security Consideration

The protection activation takes place in a controlled networking environment. Nevertheless, it is expected that the edge nodes will encapsulate and transport external traffic into separated tunnels, and the intermediate nodes will never have to process them.

9. IANA Considerations

Activation messages are encapsulated in MPLS-TP with a specific GACH channel type that needs to be assigned by IANA.

10. Normative References

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

[2] IETF draft, Multiprotocol Label Switching Transport Profile Survivability Framework (draft-ietf-mpls-tp-survive-fwk-06.txt), N. Sprecher, A. Farrel, June 2010

[3] RFC5586 - Vigoureux,, M., Bocci, M., Swallow, G., Aggarwal, R., and D. Ward, "MPLS Generic Associated Channel", May 2009.

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

[5] Crocker, D. and Overell, P.(Editors), "Augmented BNF for Syntax Specifications: ABNF", RFC 2234, Internet Mail Consortium and Demon Internet Ltd., November 1997.

11. Acknowledgments

Authors like to thank Eric Osborne, Lou Berger, Nabil Bitar and Deborah Brungard for detailed feedback on the earlier work, and the guidance and recommendation for this proposal.

We also thank Maneesh Jain, Mohit Misra, Yalin Wang, Ted Sprague, Ann Gui and Tony Jorgenson for discussion on network operation, feasibility and implementation methodology.

Authors' Addresses

Ping Pan
Email: ppan@infinera.com

Sri Mohana Satya Srinivas Singamsetty
Email: ssingamsetty@infinera.com

Rajan Rao
Email: rrao@infinera.com

Biao Lu
Email: blu@infinera.com

Sam Aldrin
Email: sam.aldrin@huawei.com

Network Working Group
Internet Draft
Expiration Date: September 2011

Y. Rekhter
Juniper Networks

R. Aggarwal
Juniper Networks

T. Morin
France Telecom

I. Grosclaude
France Telecom

N. Leymann
Deutsche Telekom AG

S. Saad
AT&T

March 14, 2011

Inter-Area P2MP Segmented LSPs

draft-raggarwa-mpls-seamless-mcast-03.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/1id-abstracts.txt.

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

Copyright and License 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.

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.

Abstract

This document describes procedures for building inter-area point-to-multipoint (P2MP) segmented service LSPs by partitioning such LSPs into intra-area segments and using BGP as the inter-area routing and label distribution protocol. Within each IGP area the intra-area segments are either carried over intra-area P2MP LSPs, using P2MP LSP hierarchy, or instantiated using ingress replication. The intra-area P2MP LSPs may be signaled using P2MP RSVP-TE or P2MP mLDP. If ingress replication is used in an IGP area then MP2P LDP LSPs or P2P RSVP-TE LSPs may be used in the IGP area. The applications/services that use such an inter-area service LSP may be BGP MVPN, VPLS multicast or Internet multicast over MPLS.

Table of Contents

1. Specification of requirements

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

2. Introduction

This document describes procedures for building inter-area point-to-multipoint (P2MP) segmented service LSPs by partitioning such LSPs into intra-area segments and using BGP as the inter-area routing and label distribution protocol. Within each IGP area the intra-area segments are either carried over intra-area P2MP LSPs, potentially using P2MP LSP hierarchy, or instantiated using ingress replication. The intra-area P2MP LSPs may be signaled using P2MP RSVP-TE or P2MP mLDP. If ingress replication is used in an IGP area then MP2P LDP or P2P RSVP-TE LSPs may be used in the IGP area. The applications/services that use such an inter-area service LSP may be BGP MVPN, VPLS multicast or Internet multicast over MPLS.

The primary use case of such segmented P2MP service LSPs is when the PEs are in different areas but in the same AS and thousands or more of PEs require P2MP connectivity. For instance this may be the case when MPLS is pushed further to the metro edge and the metros are in different IGP areas. This may also be the case when a Service Provider's network comprises multiple IGP areas in a single Autonomous System, with a large number of PEs. Seamless MPLS is the industry term to address this case [SEAMLESS-MPLS]. Thus one of the applicabilities of this document is that it describes the multicast

procedures for seamless MPLS.

It is to be noted that [BGP-MVPN], [VPLS-P2MP] already specify procedures for building segmented inter-AS P2MP service LSPs. This document complements those procedures as it extends the segmented P2MP LSP model such that it is applicable to inter-area P2MP service LSPs as well. Infact an inter-AS deployment could use inter-AS segmented P2MP LSPs as specified in [BGP-MVPN, VPLS-P2MP] where each intra-AS segment is constructed using inter-area segmented P2MP LSPs as specified in this document.

3. General Assumptions and Terminology

This document assumes BGP is used as an inter-area routing and label distribution protocol for the unicast IPv4 /32 or IPv6 /128 routes for the PEs. This document also assumes ABRs act as Route Reflectors (RR) for these routes.

Within an AS a P2MP service LSP is partitioned into 3 segments: ingress area segment, backbone area segment, and egress area segment. Within each area a segment is carried over an intra-area P2MP LSP or instantiated using ingress replication.

When intra-area P2MP LSPs are used to instantiate the intra-area segments there could be either 1:1 or n:1 mapping between intra-area segments of the inter-area P2MP service LSP and a given intra-area P2MP LSP. The latter is realized using P2MP LSP hierarchy with upstream-assigned labels [RFC5331]. For simplicity we assume that P2MP LSP hierarchy is used even with 1:1 mapping, in which case the upstream-assigned label could be an implicit NULL.

When intra-area segments of the inter-area P2MP service LSP are instantiated using ingress replication, then multiple such segments may be carried in the same P2P RSVP-TE or MP2P LDP LSP. This can be achieved using downstream-assigned labels alone.

The ingress area segment of a P2MP service LSP is rooted at a PE (or at an ASBR in the case where the P2MP service LSP spans multiple ASes). The leaves of this segment are other PEs/ASBRs and ABRs in the same area as the root PE. The backbone area segment is rooted at an ABR that is connected to the ingress area (ingress ABR), and has as its leaves ABRs that are connected to the egress area(s) or PEs in the backbone area. The egress area segment is rooted at an ABR in the egress area (egress ABR), and has as its leaves PEs and ASBR in that egress area (the latter covers the case where the P2MP service LSP spans multiple ASes). Note that for a given P2MP service LSP there may be more than one backbone segment, each rooted at its own

ingress ABR, and more than one egress area segment, each rooted at its own egress ABR.

4. Inter-area P2MP Segmented Next-Hop Extended Community

This document defines a new BGP Extended Community "Inter-area P2MP Next-Hop" extended community. This is an IP address specific Extended Community, of an extended type and is transitive across AS boundaries [RFC4360].

A PE or an ABR or an ASBR constructs the Inter-area P2MP Segmented Next-Hop Extended Community as follows:

- The Global Administrator field MUST be set to an IP address of the PE or ASBR or ABR that originates or advertises the route, which carries the P2MP Next-Hop Extended Community. For example this address may be the loopback address or the PE, ASBR or ABR that advertises the route.

- The Local Administrator field MUST be set to 0.

The detailed usage of this extended community is described in the following sections.

5. Discovering the P2MP FEC of the Inter-Area P2MP Service LSP

The P2MP FEC identifies the inter-area P2MP service LSP. The egress PEs need to learn this P2MP FEC in order to initiate the creation of the egress area segment of the P2MP inter-area service LSP.

The P2MP FEC of the inter-area P2MP LSP is learned by the egress PEs either by configuration, or based on the application-specific procedures (e.g., MVPN-specific procedures, VPLS-specific procedures).

5.1. BGP MVPN

Egress PEs discover the P2MP FEC of the service LSPs used by BGP MVPN using the I-PMSI or S-PMSI A-D routes that are originated by the ingress PEs or ASBRs following the procedures of [BGP-MVPN], along with modifications as described in this document. The NLRI of such routes encodes the P2MP FEC. The procedures in this document require that at least one ABR in a given IGP area act as Route Reflector for MVPN auto-discovery (A-D) routes.

The "Leaf Information Required" flag MUST be set in the P-Tunnel attribute carried in such routes, when originated by the ingress PEs or ASBRs. Before any Leaf auto discovery route is advertised by a PE or ABR in the same area, as described in the following sections, an I-/S-PMSI auto-discovery route is advertised either with an explicit Tunnel Type and Tunnel Identifier in the PMSI Tunnel Attribute, if the Tunnel Identifier has already been assigned, or with a special Tunnel Type of "No tunnel information present" otherwise. When the I/S-PMSI routes are re-advertised by an ABR, "Leaf Information Required" flag MUST be set in the P-Tunnel attribute present in the routes.

Note that the procedures in the above paragraph apply when intra-area segments are realized by either intra-area P2MP LSPs or by ingress replication.

When BGP MVPN I-PMSI or S-PMSI A-D routes are advertised or propagated to signal Inter-area P2MP service LSPs, they MUST carry the Inter-area P2MP Segmented Next-Hop Extended Community. This Extended Community MUST be included in the I/S-PMSI A-D route by the PE or ASBR that originates such a route and the Global Administrator field MUST be set to the advertising PE or ASBR's IP address. This Extended Community MUST also be included by ABRs as they re-advertise such routes. An ABR MUST set the Global Administrator field of the P2MP Segmented Next-Hop Extended Community to its own IP address. This allows ABRs and PEs/ASBRs to follow the procedures in this document when these procedures differ from those in [BGP-MVPN].

To avoid requiring ABRs to participate in the propagation of C-multicast routes, this document requires ABRs NOT to modify BGP Next Hop when re-advertising Inter-AS I-PMSI A-D routes. For consitancy this document requires ABRs to NOT modify BGP Next-Hop when re-advertising both Intra-AS and Inter-AS I/S-PMSI A-D routes. The egress PEs may advertise the C-multicast routes to RRs that are different than the ABRs. However ABRs still can be configured to be the Route Reflectors for C-multicast routes, in which case they will participate in the propagation of C-multicast routes.

5.2. BGP VPLS or LDP VPLS with BGP A-D

Egress PEs discover the P2MP FEC of the service LSPs used by VPLS, using the VPLS A-D routes that are originated by the ingress PEs [BGP-VPLS, VPLS-AD] or S-PMSI A-D routes that are originated by the ingress PE [VPLS-P2MP]. The NLRI of such routes encodes the P2MP FEC. The "Leaf Information Required" flag MUST be set in the P-Tunnel attribute carried in such routes. Before any Leaf auto discovery route is advertised by a PE or ABR in its own area, as described in

the following sections, an VPLS/S-PMSI autodiscovery route is advertised either with an explicit Tunnel Type and Tunnel Identifier in the PMSI Tunnel Attribute, if the Tunnel Identifier has already been assigned, or with a special Tunnel Type of "No tunnel information present" otherwise.

When VPLS A-D or S-PMSI A-D routes are advertised or propagated to signal Inter-area P2MP service LSPs, they MUST carry the Inter-area P2MP Segmented Next-Hop Extended Community. This Extended Community MUST be included in the A-D route by the PE or ASBR that originates such a route and the Global Administrator field MUST be set to the advertising PE or ASBR's IP address. This Extended Community MUST also be included by ABRs as they re-advertise such routes. An ABR MUST set the Global Administrator field of the P2MP Segmented Next-Hop Extended Community to its own IP address. This allows ABRs and PEs/ASBRs to follow the procedures in this document when these procedures differ from those in [VPLS-P2MP].

Note that the procedures in the above paragraph apply when intra-area segments are realized by either intra-area P2MP LSPs or by ingress replication.

The procedures in this document require that at least one ABR in a given area act as Route Reflector for MVPN auto-discovery (A-D) routes. These ABRs/RRs MUST NOT modify BGP Next Hop when re-advertising these A-D routes.

5.3. Internet Multicast

This section describes how the egress PEs discover the P2MP FEC when the application is internet multicast.

In the case where Internet multicast uses PIM-SM in ASM mode the following assumes that an inter-area P2MP service LSP could be used to either carry traffic on a shared (*,G), or a source (S,G) tree.

An egress PE learns the (S/*, G) of a multicast stream as a result of receiving IGMP or PIM messages on one of its IP multicast interfaces. This (S/*, G) forms the P2MP FEC of the inter-area P2MP service LSP. For each (S/*,G) for which an inter-area P2MP service LSP is instantiated, there may exist a distinct inter-area P2MP service LSP or multiple inter-area P2MP service LSPs may be aggregated using a wildcard (*, *) S-PMSI.

Note that this document does not require the use of (*, G) Inter-area P2MP service LSPs when Internet multicast uses PIM-SM in ASM mode. Infact PIM-SM in ASM mode may be supported entirely by using (S, G)

trees alone.

6. Egress PE Procedures

This section describes egress PE procedures for constructing segmented inter-area P2MP LSP. The procedures in this section apply irrespective of whether the egress PE is in a leaf IGP area, or the backbone area or even in the same IGP area as the ingress PE/ASBR.

In order to support Internet Multicast an egress PE MUST auto-configure an import Route Target with the global administrator field set to the AS of the PE and the local administrator field set to 0.

Once an egress PE discovers the P2MP FEC of an inter-area segmented P2MP service LSP, it MUST propagate this P2MP FEC in BGP in order to construct the segmented inter-area P2MP service LSP. This propagation uses BGP Leaf auto-discovery routes.

6.1. Determining the Upstream ABR/PE/ASBR

The egress PE discovers the P2MP FEC of an inter-area P2MP Segmented Service LSP as described in section 5. When an egress PE discovers this P2MP FEC it MUST first determine the upstream node to reach such a FEC. If the egress PE is in the egress area and the ingress PE is not in the that egress area, then this upstream node would be the egress ABR. If the egress PE is in the backbone area and the ingress PE is not in the backbone area, then this upstream node would be the ingress ABR. If the egress PE is in the same area as the ingress PE then this upstream node would be the ingress PE.

If the application is MVPN or VPLS then the upstream node's IP address is the IP address determined from the Global Administrator field of the Inter-area P2MP Segmented Next-hop Extended Community. As described in section 5 this Extended Community MUST be carried in the MVPN or VPLS A-D route from which the P2MP FEC of the inter-area P2MP Segmented Service LSP is determined.

If the application is Internet Multicast then the unicast routes to multicast sources/RPs SHOULD carry the VRF Route Import Extended Community [BGP-MVPN] where the IP address in the Global Administrator field is set to the IP address of the PE or ASBR advertising the unicast route. The Local Administrator field of this community MUST be set to 0. If it is not desirable to advertise the VRF Route Import Extended Community in unicast routes, then unicast routes to multicast sources/RPs MUST be advertised using the multicast SAFI

i.e. SAFI 2 and the VRF Route Import Extended Community MUST be carried in such routes.

Further if the application is internet multicast then the BGP unicast routes that advertise the route to the IP address of PEs or ASBRs or ABRs SHOULD carry the Inter-area P2MP Segmented Next-Hop Extended Community where the IP address in the Global Administrator field is set to the IP address of the PE or ASBR or ABR advertising the unicast route. The Local Administrator field of this community MUST be set to 0. If it is not desirable to advertise the P2MP Segmented Import Extended Community in BGP unicast routes, then unicast routes to ABRs, ASBRs or PEs MUST be advertised using the multicast SAFI i.e. SAFI 2 and the Inter-area P2MP Segmented Next-hop Extended Community MUST be carried in such routes. The procedures for handling the next-hop of SAFI 2 routes are the same as those of handling regular Unicast routes and follow [SEAMLESS-MPLS].

In order to determine the upstream node address the egress PE first determines the ingress PE. The egress PE determines the best route to reach S/RP. The ingress PE address is the IP address determined from the Global Administrator field of the VRF Route Import Extended Community, that is present in this route. The egress PE now finds the best unicast route to reach the ingress PE. The upstream node address is the IP address determined from the Global Administrator field of the Inter-area P2MP Segmented Next-Hop Extended Community, that is present in this route.

6.2. Originating a Leaf Auto-Discovery Route

If the P2MP FEC was derived from a MVPN or VPLS A-D route then the egress PE MUST originate a Leaf auto-discovery (A-D) route if the MVPN or VPLS A-D route carries a P-Tunnel Attribute with the "Leaf Information Required" flag set.

If the P2MP FEC was derived from an Internet Multicast S/*, G and the upstream node's address is not the same as the egress PE, then the egress PE MUST originate a Leaf auto-discovery (A-D) route.

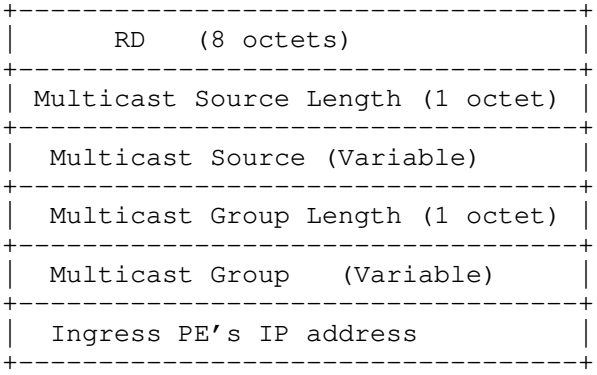
6.2.1. Leaf A-D Route for MVPN and VPLS

If the P2MP FEC was derived from MVPN or VPLS A-D routes then the Route Key field of the Leaf A-D route contains the NLRI of the A-D route from which the P2MP FEC was derived. This follows procedures for constructing Leaf A-D routes described in [BGP-MVPN, VPLS-P2MP].

6.2.2. Leaf A-D Route for Internet Multicast

If the application is internet multicast then the MCAST-VPN NLRI of the Leaf A-D route is constructed as follows:

The Route Key field of MCAST-VPN NLRI has the following format:

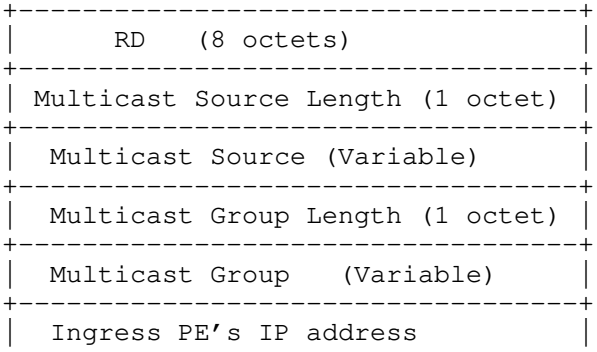


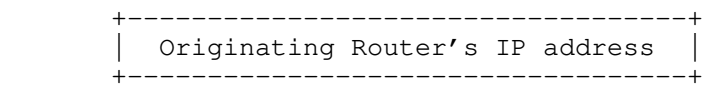
RD is set to 0 for (S,G) state and all 1s for (*,G) state, Multicast Source is set to S for (S,G) state or RP for (*,G) state, Multicast Group is set to G, Multicast Source Length and Multicast Group Length is set to either 4 or 16 (depending on whether S/RP and G are IPv4 or IPv6 addresses).

The Ingress PE's IP address is determined as described in the section "Determining the Upstream ABR/PE/ASBR".

The Originating Router's IP address field of MCAST-VPN NLRI is set to the address of the local PE (PE that originates the route).

Thus the entire MCAST-VPN NLRI of the route has the following format:





When the PE deletes (S,G)/(*,G) state that was created as a result of receiving PIM or IGMP messages on one of its IP multicast interfaces, if the PE previousely originated a Leaf auto-discovery route for that state, then the PE SHOULD withdraw that route.

6.2.3. Constructing the Rest of the Leaf A-D Route

The Next Hop field of the MP_REACH_NLRI attribute of the route SHOULD be set to the same IP address as the one carried in the Originating Router's IP Address field of the route.

When Ingress Replication is used to instantiate the egress area segment then the Leaf A-D route MUST carry a downstream assigned label in the P-Tunnel Attribute where the P-Tunnel type is set to Ingress Replication. A PE MUST assign a distinct MPLS label for each Leaf A-D route originated by the PE.

To constrain distribution of this route, the originating PE constructs an IP-based Route Target community by placing the IP address of the upstream node in the Global Administrator field of the community, with the Local Administrator field of this community set to 0. The originating PE then adds this Route Target Extended Community to this Leaf auto-discovery route. The upstream node's address is as determined in section 6.1.

The PE then advertises this route to the upstream node.

6.3. PIM-SM in ASM mode for Internet Multicast

This specification allows two options for supporting Internet Multicast with PIM-SM in ASM mode. The first option does not transit IP multicast shared trees over the MPLS network. The second option does transit shared trees over the MPLS network and relies on shared tree to source tree switchover.

6.3.1. Option 1

This option does not transit IP multicast shared trees over the MPLS network. Therefore, when an (egress) PE creates (*, G) state (as a result of receiving PIM messages on one of its IP multicast interfaces), the PE does not propagate this state using Leaf A-D

routes.

6.3.1.1. Originating Source Active auto-discovery routes

Whenever as a result of receiving PIM Register or MSDP messages an RP discovers a new multicast source the RP SHOULD originate a BGP Source Active auto-discovery route. Similarly whenever as a result of receiving MSDP messages a PE, that is not configured as a RP, discovers a new multicast source the PE SHOULD originate a BGP Source Active auto-discovery route. The BGP Source Active auto-discovery route carries a single MCAST-VPN NLRI constructed as follows:

+ The RD in this NLRI is set to 0.

+ The Multicast Source field MUST be set to S. The Multicast Source Length field is set appropriately to reflect this.

+ The Multicast Group field MUST be set to G. The Multicast Group Length field is set appropriately to reflect this.

To constrain distribution of the Source Active auto-discovery route to the AS of the advertising RP this route SHOULD carry the NO_EXPORT Community ([RFC1997]).

Using the normal BGP procedures the Source Active auto-discovery route is propagated to all other PEs within the AS.

Whenever the RP discovers that the source is no longer active, the RP MUST withdraw the Source Active auto-discovery route, if such a route was previousely advertised by the RP.

6.3.1.2. Receiving BGP Source Active auto-discovery route by PE

When as a result of receiving PIM messages on one of its IP multicast interfaces an (egress) PE creates in its Tree Information Base (TIB) a new (*, G) entry with a non-empty outgoing interface list that contains one or more IP multicast interfaces, the PE MUST check if it has any Source Active auto-discovery routes for that G. If there is such a route, S of that route is reachable via an MPLS interface, and the PE does not have (S, G) state in its TIB for (S, G) carried in the route, then the PE originates a Leaf A-D routes carrying that (S, G), as specified in Section "Leaf A-D Route for Internet Multicast".

When an (egress) PE receives a new Source Active auto-discovery route, the PE MUST check if its TIB contains an (*, G) entry with the same G as carried in the Source Active auto-discovery route. If such

an entry is found, S is reachable via an MPLS interface, and the PE does not have (S, G) state in its TIB for (S, G) carried in the route, then the PE originates a Leaf A-D routes carrying that (S, G), as specified in Section "Leaf A-D Route for Internet Multicast".

6.3.1.3. Handling (S, G, RPTbit) state

Creation and deletion of (S, G, RPTbit) state on a PE that resulted from receiving PIM messages on one of its IP multicast interfaces does not result in any BGP actions by the PE.

6.3.2. Option 2

This option does transit IP multicast shared trees over the MPLS network. Therefore, when an (egress) PE creates (*, G) state (as a result of receiving PIM messages on one of its IP multicast interfaces), the PE does propagate this state using Leaf A-D routes.

6.3.2.1. Originating Source Active auto-discovery routes

Whenever a PE creates an (S, G) state as a result of receiving Leaf A-D routes associated with Internet multicast service, if S is reachable via one of the IP multicast capable interfaces, and the PE determines that G is in the PIM-SM in ASM mode range, the PE MUST originate a BGP Source Active auto-discovery route. The route carries a single MCAST-VPN NLRI constructed as follows:

+ The RD in this NLRI is set to 0.

+ The Multicast Source field MUST be set to S. The Multicast Source Length field is set appropriately to reflect this.

+ The Multicast Group field MUST be set to G. The Multicast Group Length field is set appropriately to reflect this.

To constrain distribution of the Source Active auto-discovery route to the AS of the advertising PE this route SHOULD carry the NO_EXPORT Community ([RFC1997]).

Using the normal BGP procedures the Source Active auto-discovery route is propagated to all other PEs within the AS.

Whenever the PE deletes the (S, G) state that was previously created as a result of receiving a Leaf A-D route for (S, G), the PE that deletes the state MUST also withdraw the Source Active auto-discovery

route, if such a route was advertised when the state was created.

6.3.2.2. Receiving BGP Source Active auto-discovery route

Procedures for receiving BGP Source Active auto-discovery routes are the same as with Option 1.

6.3.2.3. Pruning Sources off the Shared Tree

If after receiving a new Source Active auto-discovery route for (S,G) a PE determines that (a) it has the (*, G) entry in its TIB, (b) the incoming interface list (iif) for that entry contains one of the IP interfaces, (c) a MPLS LSP is in the outgoing interface list (oif) for that entry, and (d) the PE does not originate a Leaf A-D route for (S,G), then the PE MUST transition the (S,G,rpt) downstream state to the Prune state. [Conceptually the PIM state machine on the PE will act "as if" it had received Prune(S,G,Rpt) from some other PE, without actually having received one.] Depending on the (S,G,rpt) state on the iifs, this may result in the PE using PIM procedures to prune S off the Shared (*,G) tree.

Transitioning the state machine to the Prune state SHOULD be done after a delay that is controlled by a timer. The value of the timer MUST be configurable. The purpose of this timer is to ensure that S is not pruned off the shared tree until all PEs have had time to receive the Source Active A-D route for (S,G).

The PE MUST keep the (S,G,rpt) downstream state machine in the Prune state for as long as (a) the outgoing interface list (oif) for (*, G) contains a MPLS LSP, and (b) the PE has at least one Source Active auto-discovery route for (S,G), and (c) the PE does not originate the Leaf A-D route for (S,G). Once either of these conditions become no longer valid, the PE MUST transition the (S,G,rpt) downstream state machine to the NoInfo state.

Note that except for the scenario described in the first paragraph of this section, in all other scenarios relying solely on PIM procedures on the PE is sufficient to ensure the correct behavior when pruning sources off the shared tree.

6.3.2.4. More on handling (S, G, RPTbit) state

Creation and deletion of (S, G, RPTbit) state on a PE that resulted from receiving PIM messages on one of its IP multicast interfaces does not result in any BGP actions by the PE.

7. Egress ABR Procedures

This section describes Egress ABR Procedures for constructing segmented inter-area P2MP LSP.

When an egress ABR receives a Leaf auto-discovery route and the Route Target extended community carried by the route contains the IP address of this ABR, then the following procedures will be executed.

If the RD of the received A-D route is not set to all 0s or all 1s, then the egress ABR MUST find a S-PMSI or I-PMSI route whose NLRI has the same value as the Route Key field of the received Leaf A-D route. If such a matching route is found then the Leaf A-D route MUST be accepted else it MUST be discarded. If the Leaf A-D route is accepted and if its the first Leaf A-D route update for the Route Key field in the route or the withdrawl of the last Leaf A-D route for the Route Key field then the following procedures will be executed.

If the RD of the received A-D route is set to all 0s or all 1s then the received Leaf A-D route is for Internet Multicast. In that case for the following procedure the Route Prefix is set to all fields of the Route Key minus the Ingress PE address. If this is the first Leaf A-D route update for this Route Prefix or the withdrawl of the last Leaf A-D route for the Route Prefix then the following procedures will be executed.

While generating a Leaf A-D route update, the egress ABR originates a Leaf A-D route, whose MCAST-VPN NLRI is constructed as follows.

The Route Key field of MCAST-VPN NLRI is the same as the Route Key field of MCAST-VPN NLRI of the received Leaf A-D route. The Originating Router's IP address field of MCAST-VPN NLRI is set to the address of the local ABR (the ABR that originates the route). In

The Next Hop field of the MP_REACH_NLRI attribute of the route SHOULD be set to the same IP address as the one carried in the Originating Router's IP Address field of the route.

To constrain distribution of this route the originating egress ABR constructs an IP-based Route Target community by placing the IP address of the upstream node in the Global Administrator field of the community, with the Local Administrator field of this community set to 0, and sets the Extended Communities attribute of this Leaf auto-discovery route to that community.

The upstream node's IP address is the IP address determined from the Global Administrator field of the Inter-area P2MP Segmented Next-hop Extended Community, where this Extended Community is obtained as

follows. When the Leaf A-D route is for MVPN or VPLS then this Extended Community is the one included in the I-S/PMSI A-D route that matches the Leaf A-D route. When the Leaf A-D route is for Internet Multicast then this Extended Community is obtained from the best unicast route to the Ingress PE. The Ingress PE address is determined from the received Leaf A-D route. The best unicast route MUST first be determined from multicast SAFI i.e., SAFI 2 routes, if present.

The ABR then advertises this Leaf A-D route to the upstream node in the backbone area.

Mechanisms specific in RFC4684 for constrained BGP route distribution can be used along with this specification to ensure that only the needed PE/ABR will have to process a said Leaf auto-discovery route.

When Ingress Replication is used to instantiate the backbone area segment then the Leaf A-D route originated by the egress ABR MUST carry a downstream assigned label in the P-Tunnel Attribute where the P-Tunnel type is set to Ingress Replication. An ABR MUST assign a distinct MPLS label for each Leaf A-D route originated by the ABR.

In order to support Internet Multicast an egress ABR MUST auto-configure an import Route Target with the global administrator field set to the AS of the ABR and the local administrator field set to 0.

When the Leaf A-D route is for Internet Multicast and if the following conditions hold true:

- Its not the first Leaf A-D route for the Route Prefix, where the Route Prefix is determined as described above

- The set of ingress PEs associated with the Route Prefix changes as a result of the new Leaf A-D route.

- The ABR determines based on local policy to propagate the Leaf A-D route towards a different ingress PE than the one to which the Leaf A-D route is being currently propagated.

Then the egress ABR MUST originate the Leaf A-D route as described in this section.

If the received Leaf A-D route is the last Leaf A-D route for the Route Key for MVPN or VPLS or for the Route Prefix, as described above, for Internet Multicast, then the ABR must withdraw the

previously advertised Leaf A-D route.

7.1. P2MP LSP as the Intra-Area LSP in the Egress Area

This section describes procedures for using intra-area P2MP LSPs in the egress area. The procedures that are common to both P2MP RSVP-TE and P2MP LDP are described first, followed by procedures that are specific to the signaling protocol.

When P2MP LSPs are used as the intra-area LSPs, note that an existing intra-area P2MP LSP may be used solely for a particular inter-area P2MP service LSP, or for other inter-area P2MP service LSPs as well. The choice between the two options is purely local to the egress ABR. The first option provides one-to-one mapping between inter-area P2MP service LSPs and intra-area P2MP LSPs; the second option provides many-to-one mapping, thus allowing to aggregate forwarding state.

7.1.1. RD of the received Leaf-AD route is not zero or all ones

When the RD of the received Leaf A-D route is not set to zero or all ones then the ABR MUST re-advertise in the egress area the MVPN/VPLS A-D route, that matches the Leaf A-D route to signal the binding of the intra-area P2MP LSP to the inter-area P2MP service LSP. This must be done ONLY if a) such a binding hasn't already been advertised or b) The binding has changed. The re-advertised route MUST carry the Inter-area P2MP Segmented Next-Hop Extended Community.

The PMSI Tunnel attribute of the re-advertised route specifies either an intra-area P2MP RSVP-TE LSP or an intra-area P2MP LDP LSP rooted at the ABR and MUST also carry an upstream assigned MPLS label. The upstream-assigned MPLS label MUST be set to implicit NULL if the mapping between the inter-area P2MP service LSP and the intra-area P2MP LSP is one-to-one. If the mapping is many-to-one the intra-area segment of the inter-area P2MP service LSP (referred to as the "inner" P2MP LSP) is constructed by nesting the inter-area P2MP service LSP in an intra-area P2MP LSP (referred to as the "outer" intra-area P2MP LSP), by using P2MP LSP hierarchy based on upstream-assigned MPLS labels [RFC 5332].

If segments of multiple MVPN or VPLS S-PMSI service LSPs are carried over a given intra-area P2MP LSP, each of these segments MUST carry a distinct upstream-assigned label, even if all these service LSPs are for (C-S/*, C-G/*)s from the same MVPN/VPLS. Therefore, an ABR maintains an LFIB state for each of the (C-S/*, C-G/*)s carried over S-PMSIs traversting this ABR (that applies to both the ingress and

the egress ABRs).

7.1.2. RD of the received Leaf A-D route is zero or all ones

When the RD of the received Leaf A-D route is set to zero or all ones then this is the case of inter-area P2MP service LSP being associated with the Internet multicast service. The procedures for this are described below.

7.1.2.1. Internet Multicast and S-PMSI A-D Routes

This section applies only if it is desirable to send a particular Internet Multicast flow to only those egress PEs that have receivers in a particular (S, G) or a particular (*, G) multicast flow.

The egress ABR MUST originate a S-PMSI A-D route. The PMSI Tunnel attribute of the route MUST contain the identity of the intra-area P2MP LSP and an upstream assigned MPLS label. The RD, Multicast Source Length, Multicast Source, Multicast Group Length (1 octet), and Multicast Group fields of the NLRI of this route are the same as of the Leaf A-D route. The egress ABR MUST advertise this route into the backbone area. The Route Target of this route is an AS specific route-target with the AS set to the AS of the advertising ABR while the local administrator field is set to 0.

7.1.2.2. Internet Multicast and Wildcard S-PMSI A-D Routes

It may be desirable for an ingress PE to aggregate Internet Multicast routes over a single Inter-area P2MP LSP. This can be achieved using wildcard, i.e., (*,*) S-PMSI A-D routes. An ingress PE MAY advertise a wildcard S-PMSI route as described in section "Ingress PE Procedures". If the ingress PE does indeed originate such a route the egress ABR would receive this route from the ingress ABR and MUST re-advertise it with the PMSI Tunnel Attribute containing the identifier of the intra-area P2MP LSP in the egress area and an upstream assigned label assigned to the inter-area wildcard S-PMSI.

7.1.3. Internet Multicast and the Expected Upstream Node

If the mapping between the inter-area P2MP service LSP for Internet multicast service and the intra-area P2MP LSP is many-to-one then an egress PE must be able to determine whether a given multicast packet for a particular (S, G) is received from the "expected" upstream node. The expected node is the node towards which the Leaf A-D route

is sent by the egress PE. Packets received from another upstream node for that (S, G) MUST be dropped. To allow the egress PE to determine the sender upstream node, the intra-area P2MP LSP must be signaled with no PHP, when the mapping between the inter-area P2MP service LSP for Internet multicast service and the intra-area P2MP LSP is many-to-one.

Further the egress ABR MUST first push onto the label stack the upstream assigned label advertised in the S-PMSI route, if the label is not an Implicit NULL.

7.1.4. P2MP LDP LSP as the Intra-Area P2MP LSP in the Egress Area

The procedures above are sufficient if P2MP LDP LSPs are used as the Intra-area P2MP LSP in the Egress area.

7.1.5. P2MP RSVP-TE LSP as the Intra-Area P2MP LSP in the Egress Area

If P2MP RSVP-TE LSP is used as the the intra-area LSP in the egress area, then the egress ABR can either (a) graft the leaf (whose IP address is specified in the received Leaf auto-discovery route) into an existing P2MP LSP rooted at the egress ABR, and use that LSP for carrying traffic for the inter-area segmented P2MP service LSP, or (b) originate a new P2MP LSP to be used for carrying (S,G).

When the RD of the received Leaf A-D route is zero or all ones, then the procedures are as described in section 7.1.2 ("RD of the received Leaf A-D route is zero or all ones").

Note also that the SESSION object that the egress ABR would use for the intra-area P2MP LSP need not encode the P2MP FEC from the received Leaf auto-discovery route.

7.2. Ingress Replication in the Egress Area

When Ingress Replication is used to instantiate the egress area segment then the Leaf A-D route advertised by the egress PE MUST carry a downstream assigned label in the P-Tunnel Attribute where the P-Tunnel type is set to Ingress Replication. We will call this the egress PE downstream assigned label.

The egress ABR MUST forward packets received from the backbone area intra-area segment, for a particular inter-area P2MP LSP, to all the egress PEs from which the egress ABR has imported a Leaf A-D route for the inter-area P2MP LSP. A packet to a particular egress PE is

encapsulated, by the egress ABR, using a MPLS label stack the bottom label of which is the egress PE downstream assigned label. The top label is the P2P RSVP-TE or the MP2P LDP label to reach the egress PE.

Note that these procedures ensures that an egress PE always receives packets only from the expected upstream PE.

8. Ingress ABR Procedures for constructing segmented inter-area P2MP LSP

When an ingress ABR receives a Leaf auto-discovery route and the Route Target extended community carried by the route contains the IP address of this ABR, then the following procedures will be executed.

These procedures are the same as in the section "Egress ABR Procedures" with egress ABR replaced with ingress ABR, backbone area replaced with ingress area and backbone area segment replaced with ingress area segment.

In order to support Internet Multicast the ingress ABR MUST auto-configure an import Route Target with the global administrator field set to the AS of the ABR and the local administrator field set to 0.

8.1. P2MP LSP as the Intra-Area LSP in the Backbone Area

If the RD of the received Leaf A-D route is not zero, and P2MP LSP is used as the the intra-area LSP in the backbone area, then the procedures for binding the backbone area segment of the inter-area P2MP LSP to the intra-area P2MP LSP in the backbone area, are the same as in section "Egress ABR Procedures" and sub-section "P2MP LSP as the Intra-Area LSP in the Egress Area".

When the RD of the received Leaf A-D route is zero, as is the case where the inter-area service P2MP LSP is associated with the Internet multicast service, then the procedures are the same as in section "Egress ABR Procedures", and and sub-section "P2MP LSP as the Intra-Area LSP in the Egress Area", with egress ABR replaced with the ingress ABR. It is to be noted that if the backbone area uses wildcard S-PMSI then the egress area also must use wildcard S-PMSI for Internet Multicast or the ABRs must merge the wildcard S-PMSI onto the egress area (S, G) or (*, G) S-PMSI. The procedures for such merge require IP processing on the ABRs.

8.2. Ingress Replication in the Backbone Area

When Ingress Replication is used to instantiate the backbone area segment then the Leaf A-D route advertised by the egress ABR MUST carry a downstream assigned label in the P-Tunnel Attribute where the P-Tunnel type is set to Ingress Replication. We will call this the egress ABR downstream assigned label. The egress ABR MUST assign a distinct MPLS label for each Leaf A-D route originated by the ABR.

The ingress ABR MUST forward packets received from the ingress area intra-area segment, for a particular inter-area P2MP LSP, to all the egress ABRs from which the ingress ABR has imported a Leaf A-D route for the inter-area P2MP LSP. A packet to a particular egress ABR is encapsulated, by the inress ABR, using a MPLS label stack the bottom label of which is the egress ABR downstream assigned label. The top label is the P2P RSVP-TE or the MP2P LDP label to reach the egress ABR.

9. Ingress PE/ASBR Procedures

This section describes Ingress PE/ASBR procedures for constructing segmented inter-area P2MP LSP.

When an ingress PE/ASBR receives a Leaf auto-discovery route and the Route Target extended community carried by the route contains the IP address of this PE/ASBR, then the following procedures will be executed.

If the RD of the received A-D route is not set to all 0s or all 1s, then the egress ABR MUST find a S-PMSI or I-PMSI route whose NLRI has the same value as the Route Key field of the received Leaf A-D route. If such a matching route is found then the Leaf A-D route MUST be accepted else it MUST be discarded. If the Leaf A-D route is accepted then it MUST be processed as per MVPN or VPLS procedures.

If the RD of the received A-D route is set to all 0s or all 1s then the received Leaf A-D route is for Internet Multicast. In that case for the following procedure the Route Prefix is set to all fields of the Route Key minus the Ingress PE address. If this is the first Leaf A-D route update for this Route Prefix or the withdrawl of the last Leaf A-D route for the Route Prefix then the following procedures will be executed. The information carried in the MCAST-VPN NLRI of the route MUST be decoded. The PIM implementation should set its upstream (S/RP,G) state machine in Joined state for the (S/RP, G) received via a Leaf auto-discovery route update. Likewise, the PIM implementation should set its upstream (S/RP, G) state machine in Pruned state for the (S/RP, G) received via a Leaf auto-discovery

route withdrawl.

9.1. P2MP LSP as the intra-area LSP in the ingress area

If the RD of the received Leaf A-D route is not zero, and P2MP LSP is used as the the intra-area LSP in the ingress area, then the procedures for binding the ingress area segment of the inter-area P2MP LSP to the intra-area P2MP LSP in the ingress area, are the same as in section "Egress ABR Procedures" and sub-section "P2MP LSP as the Intra-Area LSP in the Egress Area".

When the RD of the received Leaf A-D route is zero, as is the case where the inter-area service P2MP LSP is associated with the Internet multicast service, then the ingress PE may originate a S-PMSI route with the RD, multicast source, multicast group fields being the same as those in the received Leaf A-D route.

Further an ingress PE may originate a wildcard S-PMSI route as per the procedures in [MVPN-WILDCARD-SPMSI] with the RD set to 0. This route may be originated by the ingress PE based on configuration or based on the import of a Leaf A-D route with RD set to 0. If an ingress PE originates such a route, then the ingress PE may decide not to originate (S, G) or (*, G) S-PMSI routes.

It is to be noted that if ingress area uses wildcard S-PMSI then the backbone area also must use wildcard S-PMSI for Internet Multicast or the ABRs must merge the wildcard S-PMSI onto the backbone area (S, G) or (*, G) S-PMSI. The procedures for such merge require IP processing on the ABRs.

9.2. Ingress Replication in the Ingress Area

When Ingress Replication is used to instantiate the ingress area segment then the Leaf A-D route advertised by the ingress ABR MUST carry a downstream assigned label in the P-Tunnel Attribute where the P-Tunnel type is set to Ingress Replication. We will call this the ingress ABR downstream assigned label. The ingress ABR MUST assign a distinct MPLS label for each Leaf A-D route originated by the ABR.

The ingress PE/ASBR MUST forward packets received from the CE, for a particular inter-area P2MP LSP, to all the ingress ABRs from which the ingress PE/ASBR has imported a Leaf A-D route for the inter-area P2MP LSP. A packet to a particular ingress ABR is encapsulated, by the inress PE/ASBR, using a MPLS label stack the bottom label of which is the ingress ABR downstream assigned label. The top label is the P2P RSVP-TE or the MP2P LDP label to reach the ingress ABR.

10. Common Tunnel Type in the Ingress and Egress Areas

For a given inter-area service P2MP LSP, the PE/ASBR that is the root of that LSP controls the tunnel type of the intra-area P-tunnel that carries the ingress area segment of that LSP. However, the tunnel type of the intra-area P-tunnel that carries the backbone area segment of that LSP may be different from the tunnel type of the intra-area P-tunnels that carry the ingress area segment and the egress area segment of that LSP. In that situation if for a given inter-area P2MP LSP it is desirable/necessary to use the same tunnel type for the intra-area P-tunnels that carry the ingress area segment and the egress area segment of that LSP, then the following procedures on the ingress ABR and egress ABR provide this functionality.

When an ingress ABR re-advertises into the backbone area a BGP MVPN I-PMSI, or S-PMSI A-D route, or VPLS A-D route, the ingress ABR places the PMSI Tunnel attribute of this route into the ATTR_SET BGP Attribute [L3VPN-IBGP], adds this attribute to the re-advertised route, and then replaces the original PMSI Tunnel attribute with a new one (note, that the Tunnel type of the new attribute may be different from the Tunnel type of the original attribute).

When an egress ABR re-advertises into the egress area a BGP MVPN I-PMSI or S-PMSI A-D route, or VPLS A-D route, if the route carries the ATTR_SET BGP attribute [L3VPN-IBGP], then the ABR sets the Tunnel type of the PMSI Tunnel attribute in the re-advertised route to the Tunnel type of the PMSI Tunnel attribute carried in the ATTR_SET BGP attribute, and removes the ATTR_SET from the route.

11. Placement of Ingress and Egress PEs

As described in earlier sections, procedures in this document allow the placement of ingress and egress PEs in the backbone area. They also allow the placement of egress PEs in the ingress area or the placement of ingress PEs in the egress area.

For instance ABRs in the backbone area may act as ingress and egress PEs for Internet Multicast, as per the ingress and egress PE definition in this document. This may be the case if the service is Internet Multicast and relies on Internet Multicast in the ingress and egress areas and its desirable to carry Internet Multicast over MPLS in the backbone area. This may also be the case if the service is Multicast VPN and the P-tunnel technology in the ingress and egress areas uses PIM based IP/GRE P-tunnels. As far as the ABRs are concerned PIM signaling for such P-Tunnels is handled as per the ingress/egress PE Internet Multicast procedures in this document. To

facilitate this the ABRs may advertise their loopback addresses in BGP using multicast-SAFI i.e., SAFI 2, if non-congruence between unicast and multicast is desired.

12. Data Plane

This section describes the data plane procedures on the ABRs, ingress PEs, egress PEs and transit routers.

12.1. Data Plane Procedures on an ABR

When procedures in this document are followed to signal inter-area P2MP Segmented LSPs then ABRs are required to perform only MPLS switching. When an ABR receives a MPLS packet from an "incoming" intra-area segment of the inter-area P2MP Segmented LSP, it forwards the packet, based on MPLS switching, onto another "outgoing" intra-area segment of the inter-area P2MP Segmented LSP.

If the outgoing intra-area segment is instantiated using a P2MP LSP, and if there is a one-to-one mapping between the outgoing intra-area segment and the P2MP LSP, then the ABR MUST pop the incoming segment's label stack and push the label stack of the outgoing P2MP LSP. If there is a many-to-one mapping between outgoing intra-area segments and the P2MP LSP then the ABR MUST pop the incoming segment's label stack and first push the upstream assigned label corresponding to the outgoing intra-area segment, if such a label has been assigned, and then push the label stack of the outgoing P2MP LSP.

If the outgoing intra-area segment is instantiated using ingress replication then the ABR must pop the incoming segment's label stack and replicate the packet once to each leaf ABR or PE of the outgoing intra-area segment. The label stack of the packet sent to each such leaf MUST first include a downstream assigned label assigned by the leaf to the segment, followed by the label stack of the P2P or MP2P LSP to the leaf.

12.2. Data Plane Procedures on an Egress PE

An egress PE must first identify the inter-area P2MP segmented LSP based on the incoming label stack. After this identification the egress PE must forward the packet using the application that is bound to the inter-area P2MP segmented LSP.

Note that the application specific forwarding for MVPN service may

require the egress PE to determine whether the packets were received from the expected sender PE. When the application is MVPN then the FEC of an inter-area P2MP Segmented LSP is at the granularity of the sender PE. Note that MVPN intra-AS I-PMSI A-D routes and S-PMSI A-D routes both carry the Originating Router IP Address. Thus an egress PE could associate the data arriving on P-tunnels advertised by these routes with the Originating Router IP Address carried by these routes which is the same as the ingress PE. Since a unique label stack is associated with each such FEC, the egress PE can determine the sender PE from the label stack.

Likewise for VPLS service for the purposes of MAC learning the egress PE must be able to determine the "VE-ID" from which the packets have been received. The FEC of the VPLS A-D routes carries the VE-ID. Thus an egress PE could associate the data arriving on P-tunnels advertised by these routes with the VE-ID carried by these routes. Since a unique label stack is associated with each such FEC, the egress PE can perform MAC learning for packets received from a given VE-ID.

When the application is Internet Multicast it is sufficient for the label stack to include identification of the sender upstream node. When P2MP LSPs are used this requires that PHP MUST be turned off. When Ingress Replication is used the egress PE knows the incoming downstream assigned label to which it has bound a particlar (S/*, G) and must accept packets with only that label for that (S/*. G).

12.3. Data Plane Procedures on an Ingress PE

The Ingress PE must perform application specific forwarding procedures to identify the outgoing inta-area segment of an incoming packet.

If the outgoing intra-area segment is instantiated using a P2MP LSP, and if there is a one-to-one mapping between the outgoing intra-area segment and the P2MP LSP, then the ingress PE MUST encapsulate the packet in the label stack of the outgoing P2MP LSP. If there is a many-to-one mapping between outgoing intra-area segments and the P2MP LSP then the PE MUST first push the upstream assigned label corresponding to the outgoing intra-area segment, if such a label has been assigned,
 and then push the label stack of the outgoing P2MP LSP.

If the outgoing intra-area segment is instantiated using ingress replication then the PE must replicate the packet once to each leaf ABR or PE of the outgoing intra-area segment. The label stack of the packet sent to each such leaf MUST first include a downstream

assigned label assigned by the leaf to the segment, followed by the label stack of the P2P or MP2P LSP to the leaf.

12.4. Data Plane Procedures on Transit Routers

When procedures in this document are followed to signal inter-area P2MP Segmented LSPs then tansit routers in each area perform only MPLS switching.

13. IANA Considerations

This document defines a new BGP Extended Community called "Inter-area P2MP Segmented Next-Hop". This community is IP Address Specific, of an extended type, and is transitive. A codepoint for this community should be assigned both from the IPv4 Address Specific Extended Community registry, and from the IPv6 Address Specific Extended Community registry. The same code point should be assigned from both registries.

14. Security Considerations

These will be spelled out in a future revision.

15. References

15.1. Normative References

[RFC5332] T. Eckert, E. Rosen, R. Aggarwal, Y. Rekhter, RFC5332

[RFC2119] "Key words for use in RFCs to Indicate Requirement Levels.", Bradner, March 1997

[MVPN-BGP] "BGP Encodings and Procedures for Multicast in MPLS/BGP IP VPNs", R. Aggarwal, E. Rosen, T. Morin, Y. Rekhter, draft-ietf-l3vpn-2547bis-mcast-bgp

[[VPLS-P2MP] "Multicast in VPLS", R. Aggarwal, Y. Kamite, L. Fang, draft-ietf-l2vpn-vpls-mcast

[L3VPN-IBGP] "Internal BGP as PE-CE protocol", Pedro Marques, et al., draft-ietf-l3vpn-ibgp, work in progress

15.2. Informative References

[SEAMLESS-MPLS] "Seamless MPLS Architecture", N. Leymann et. al., draft-leymann-mpls-seamless-mpls

16. Author's Address

Yakov Rekhter
Juniper Networks
1194 North Mathilda Ave.
Sunnyvale, CA 94089
Email: yakov@juniper.net

Rahul Aggarwal
Juniper Networks
1194 North Mathilda Ave.
Sunnyvale, CA 94089
Phone: +1-408-936-2720
Email: rahul@juniper.net

Thomas Morin
France Telecom R & D
2, avenue Pierre-Marzin
22307 Lannion Cedex
France
Email: thomas.morin@francetelecom.com

Irene Grosclaude
France Telecom R & D
2, avenue Pierre-Marzin
22307 Lannion Cedex
France
Email: irene.grosclaude@orange-ftgroup.com

Nicolai Leymann
Deutsche Telekom AG
Winterfeldtstrasse 21
Berlin 10781
DE
Email: n.leymann@telekom.de

Samir Saad
AT&T
Email: ss2539@att.com

Network Working Group
Internet Draft
Intended status: Standards Track
Expires: April 17, 2011

Kamran Raza
Cisco Systems

Sami Boutros
Cisco Systems

October 18, 2010

LDP IP and PW Capability

draft-raza-mpls-ldp-ip-pw-capability-00.txt

Status of this Memo

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

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

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

The list of current Internet-Drafts can be accessed at http://www.ietf.org/ietf/1id-abstracts.txt

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

This Internet-Draft will expire on April 17, 2011.

Copyright Notice

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

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

publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document.

Abstract

Currently, no LDP capability is exchanged for LDP applications like IP label switching and L2VPN/PW signaling. When an LDP session comes up, an LDP speaker may unnecessarily advertise its local state for such LDP applications even when the peer session may be established for some other applications like ICCP. This document proposes a solution by which an LDP speaker announces it "incapability" or disability or non-support for IP label switching or L2VPN/PW application, hence disabling corresponding application state exchange over established LDP session.

Table of Contents

1. Introduction

LDP Capabilities [RFC5561] introduced a mechanism to negotiate LDP capabilities for given feature amongst peer LSRs. This mechanism insures that no unnecessary state is exchanged between peer LSRs unless corresponding feature capability is successfully negotiated between peers.

While new features and applications like Typed Wildcard FEC [RFC5918], Inter-Chassis Communication Protocol [ICCP], and mLDP [MLDP] make use of LDP capabilities framework for their feature negotiation, the earlier LDP features and applications like IP label switching and L2VPN/PW signaling [RFC4447] may cause unnecessary state exchange between LDP peers if the given application is not enabled on one of the LDP speakers participating in a given session. For example, when bringing up and using an LDP peer session with a remote PE LSR for purely ICCP signaling purposes, the LDP speaker may unnecessarily advertise labels for IP (unicast) prefixes to this ICCP related LDP peer as per its default behavior. To avoid this unnecessary state advertisement and exchange, currently customers are typically required to configure/define some sort of LDP state/label filtering policies on the box, which introduces operational overhead and complexity.

This document proposes a solution by which an LDP speaker announces its "incapability" (or disability) to its peer for IP Label Switching and/or L2VPN/PW Signaling application at session establishment time. This helps avoiding unnecessary state exchange for such feature applications. The proposal also allows a previously disabled application to be enabled later during the session lifetime. The document introduces two new LDP Capabilities for IP label switching and L2VPN/PW applications to implement the proposal.

2. Conventions used in this document

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

The term "IP" in this document refers to "IP unicast", unless otherwise explicitly stated.

3. Non-negotiated LDP applications

For the applications that existed before LDP Capabilities [RFC5561] mechanics were defined, LDP speaker may advertise relevant application state to its peers after session establishment without waiting for any capabilities exchange and negotiation.

The most important non-negotiated applications include:

- IP [v4 and v6] label switching
- L2VPN/PW signaling

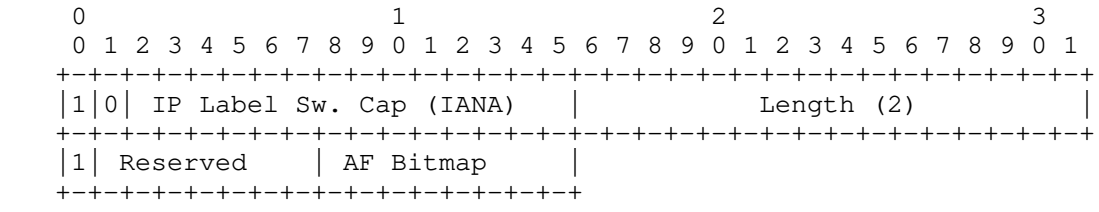
To disable unnecessary state exchange for such LDP applications, two new capabilities are being introduced in this document. These new capabilities allow an LDP speaker to notify its LDP peer at the session establishment time when one or more LDP "Non-negotiated applications" are not required/configured on the sender side. Upon receipt of such capability TLV, the receiving LDP speaker MUST disable the advertisement of application state towards the sender. These capabilities can also be sent later in a Capability message to either disable these applications, or to enable previously disabled applications.

3.1. Application Control Capabilities

To control advertisement of state related to non-negotiated LDP applications, namely IP Label switching and L2VPN/PW signaling, two new capability TLVs are defined as described in the following subsections.

3.1.1. IP Label Switching Capability TLV

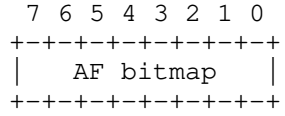
The IP Label Switching capability is a new Capability Parameter defined with the following format:



The value of the U-bit for the IP capability parameter TLV MUST be set to 1 so that a receiver MUST silently ignore this TLV if unknown

to it, and continue processing the rest of the message. Once advertised, this capability cannot be withdrawn and hence the S-bit must always be set to 1 both in Initialization message and Capability message. The capability data associated with this TLV is 1 byte long "Address Family Bitmap", and hence the TLV length MUST be set to 2.

The Capability data "Address Family Bitmap" is defined as:



Where:

bit0: IPv4 label switching application

bit1: IPv6 label switching application

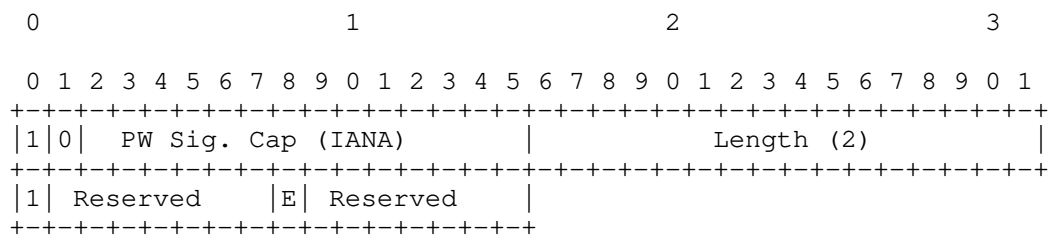
bit2-7: Reserved.

A bit in the bitmap is set to 0 or 1 to disable or enable respectively a corresponding IP application.

As described earlier, "IP Label Switching" Capability Parameter TLV MAY be included by an LDP speaker in an Initialization message to signal to its peer LSR that state exchange for IPv4 and/or IPv6 application(s) need to be disabled on a given peer session. This TLV can also be sent later in a Capability message to selectively enable or disable IPv4/v6 label switching application(s).

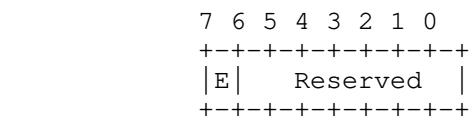
3.1.2. PW Signaling Capability TLV

The "PW Signaling" capability is a new Capability Parameter defined with the following format:



The value of the U-bit for the PW capability parameter TLV MUST be set to 1 so that a receiver MUST silently ignore this TLV if unknown to it, and continue processing the rest of the message. Once advertised, this capability cannot be withdrawn and hence the S-bit must always be set to 1 in Initialization message or Capability message. The capability data associated with this TLV is 1 byte long and hence the TLV length MUST be set to 2.

The capability data is defined as following byte:



Where E-bit (Enable bit) is used to control PW signaling application by setting it to 0 and 1 to disable and enable the application respectively.

As described earlier, PW Signaling Capability Parameter TLV MAY be included by an LDP speaker in an Initialization message to signal to its peer LSR that state exchange for PW application need to be disabled on given peer session. This TLV can also be sent later in a Capability message to selectively enable/disable the PW Signaling application.

3.2. Procedures for Application Control Capabilities in an Initialization message

LDP Capabilities [RFC5561] dictate that the S-bit of capability parameter in an Initialization message MUST be set to 1 and SHOULD be ignored on receipt.

An LDP speaker determines (e.g. via some local configuration or default policy) if they need to disable IP and/or L2VPN/PW applications with a peer LSR. If there is a need to disable, then the IP and/or PW application capability TLVs need to be included in the Initialization message with respective application bits set to 0 to indicate application disable, where the application bit refers to a bit in "Address Family Bitmap" of the "IP Label Switching" Capability or E-bit in "PW Signaling" Capability.

An LDP speaker that supports the "IP Label Switching" and/or "PW Signaling" capability MUST interpret those TLVs in a received Initialization message such that it disables the advertisement of the

application state towards the sender LSR for IP (v4 and/or v6) and/or L2VPN/PW applications if their application control bits are set to 0. If a receiving LDP speaker does not understand the capability TLVs, then it MUST respond to the sender with "Unsupported TLV" Notification as described in LDP Capabilities [RFC5561]. Upon receipt of such Notification, the sender MAY still continue to block/disable its outbound state advertisement towards the peer for the requested disabled applications.

Once this capability has been sent by sender LSR and received and understood by the receiver LSR, then both these LSRs MUST NOT exchange any state related to the disabled applications until and unless these applications are explicitly enabled again (e.g. via the same Capability TLV sent in a Capability message with corresponding application control bit set to 1).

"IP Label Switching" and "PW Signaling" capability TLVs are unilateral/uni-directional in nature. This means that the receiving LSR may not need to send a similar capability TLV in an Initialization or Capability message towards the sender. This unilateral behavior also conforms to the procedures defined in the Section 6 of LDP Capabilities [RFC 5561].

3.3. Procedures for Application Control capabilities in a Capability message

If the LDP peer supports "Dynamic Announcement Capability" [RFC5561], then an LDP speaker can send IP Label Switching and/or PW Signaling capability in a Capability message. Once advertised, these capabilities cannot be withdrawn and hence the S-bit of the TLV MUST be set to 1 when sent in a Capability message.

An LDP speaker may decide to send this TLV towards an LDP peer if any of its IP and/or L2VPN/PW signaling applications gets disabled or if previously disabled IP or L2VPN/PW application(s) gets enabled again. In this case, LDP speaker constructs the TLVs with appropriate application control bitmap and sends the corresponding capability TLVs in a Capability message. Furthermore, the LDP speaker also withdraws application(s) related advertised state (such as label bindings) from its peer.

Upon receipt of those TLVs in a Capability message, the receiving LDP speaker reacts in the same manner as it reacts upon the receipt of those TLVs in an Initialization message. Additionally, the receiving LDP speaker withdraws the application(s) related advertised state (such as label bindings) from the sending LDP speaker. If the receiving LDP speaker does not understand or support either Dynamic

Announcement capability or received Application Control capability TLV ("IP Label Switching" or "PW Signaling"), it MUST respond with "Unsupported Capability" notification to the sender of the Capability message.

4. Operational Examples

4.1. Disabling IP/PW label applications on an ICCP session

Consider two PE routers, LSR1 and LSR2, which understand/support "IP Label Switching" and "PW Signaling" capability TLVs. These LSR have an established LDP session due to ICCP application in order to exchange ICCP state related to dual-homed devices connected to these LSRs. Let us assume that LSR1 is provisioned not to exchange any label bindings related to IP (v4/v6) prefixes and PW layer2 FEC (FEC128/129) with LSR2.

To indicate its "disability" for the IP/PW applications, the LSR1 will include both the "IP Label Switching" capability TLV (with bit0-1 of "Address Family Bitmap" set to 0) and "PW Signaling" capability TLV (with E-bit set to 0) in the Initialization message. Upon receipt of those TLVs in Initialization message, the LSR2 will disable any IP/PW address/label binding state advertisement towards LSR1.

The LSR1 will also disable any IP/PW address/label binding state towards LSR2, irrespective of the fact whether or not LSR2 could disable the corresponding application state advertisement towards LSR1.

4.2. Disabling IP Label Switching application on a L2VPN/PW session

Now, consider LSR1 and LSR2 have an established session due to L2VPN/PW application in order to exchange PW (FEC128/129) label bindings for VPWS/VPLS services amongst them. Since in most typical deployments, there is no need to exchange IP (v4/v6) address/label bindings amongst the PE LSRs, let us assume that LSR1 is provisioned to disable IP (v4/v6) application on given PW session towards LSR2.

To indicate its disability for IP application, the LSR1 will include the "IP Label Switching" capability TLV in the Initialization message with bit0-1 (IPv4, IPv6) in "Address Family Bitmap" set to zero. Upon receipt of this TLV in Initialization message, the LSR2 will disable any IP address/label binding state advertisement towards LSR1.

The LSR1 will also disable any IP address/label binding state towards LSR2, irrespective of the fact whether or not LSR2 could disable the corresponding IP application state advertisement towards LSR1.

4.3. Disabling IP application dynamically on an established IP/PW session

Assume that LSRs from previous sections were initially provisioned to exchange both IP and PW state over the session between them, and also support "Dynamic Announcement" capability [RFC5561]. Now, assume that LSR1 is provisioned to disable IP label switching application with LSR2. In this case, LSR1 will first withdraw all its IP label state by sending a single Label Withdraw message with IP prefix Typed Wildcard FEC using the mechanics described in [RFC5918], and Address Withdraw message to withdraw its addresses. LSR1 will also send IP Label Switching capability TLV in Capability message towards LSR2 with bit0-1 (IPv4, IPv6) in "Address Family Bitmap" set to zero. Upon receipt of this TLV, LSR2 will also disable IP application towards LSR1 and withdraw all previous IP application label/address state using the same mechanics as described earlier for LSR1. The disability of IP label switching dynamically should not impact L2VPN/PW application on given session, and both LSRs should continue to exchange PW Signaling application related state.

5. Security Considerations

The proposal introduced in this document does not introduce any new security considerations beyond that already apply to the base LDP specification [RFC5036] and [RFC5920].

6. IANA Considerations

The document introduces following two new capability parameter TLVs and requests following LDP TLV code point assignment by IANA:

- "IP Label Switching" Capability TLV (requested codepoint: 0x50C)
- "PW Signaling" Capability TLV (requested codepoint: 0x50D)

7. Conclusions

The document proposed a solution using LDP Capabilities [RFC5561] mechanics to disable unnecessary state exchange, if/as desired, between LDP peers for currently non-negotiated IP/PW applications.

8. References

8.1. Normative References

[RFC5561] Thomas, B., Raza, K., Aggarwal, S., Aggarwal, R., and Le Roux, JL., "LDP Capabilities", RFC 5561, July 2009.

[RFC5918] Asati, R., Minei, I., and Thomas, B. "Label Distribution Protocol Typed Wildcard FEC", RFC 5918, August 2010.

[ICCP] Martini, L., Salam, S., and Matsushima, S., "Inter-Chassis Communication Protocol for L2VPN PE Redundancy", draft-ietf-pwe3-iccp-03.txt, Work in Progress, July 2010.

[MLDP] Minei, I., Kompella, K., Wijnands, I., and Thomas, B., "LDP Extensions for Point-to-Multipoint and Multipoint-to-Multipoint Label Switched Paths", draft-ietf-mpls-ldp-p2mp-10.txt, Work in Progress, July 2010.

[RFC4447] L. Martini, Editor, E. Rosen, El-Aawar, T. Smith, G. Heron, "Pseudowire Setup and Maintenance using the Label Distribution Protocol", RFC 4447, April 2006.

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

8.2. Informative References

[RFC5036] Andersson, L., Menei, I., and Thomas, B., Editors, "LDP Specification", RFC 5036, September 2007.

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

9. Acknowledgments

The authors would like to thank Eric Rosen for his valuable input and comments.

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

Authors' Addresses

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

Sami Boutros
Cisco Systems, Inc.
3750 Cisco Way,
San Jose, CA 95134, USA.
E-mail: sboutros@cisco.com

MPLS
Internet-Draft
Intended status: Informational
Expires: April 5, 2013

C. Villamizar, Ed.
Outer Cape Cod Network
Consulting
October 2, 2012

Use of Multipath with MPLS-TP and MPLS
draft-villamizar-mpls-tp-multipath-03

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 can LSP can be carried over multipath links while also providing a fully MPLS-TP compliant server layer for MPLS-TP LSP. This document describes the means of supporting MPLS as a server layer for MPLS-TP. The use of MPLS-TP LSP as a server layer for MPLS LSP 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 April 5, 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

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 LSP. Some vendors support load split across equal cost MPLS LSP where the load is split proportionally to the reserved bandwidth of the set of LSP.

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 an MPLS-TP LSP due to the inherent MPLS-TP capacity limitation imposed by MPLS-TP OAM packet ordering constraints.

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 LSP 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 acheive 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

MPLS LSP may be used as a server layer for MPLS-TP LSP as long as all MPLS-TP requirements are met, including the requirement that packets within an MPLS-TP LSP are not reordered, including both payload and OAM packets.

Supporting MPLS-TP LSP overa fully MPLS-TP conformant MPLS LSP server layer where the MPLS LSP are making use of multipath, requires special treatment of the MPLS-TP LSP such that those LSP only are not subject to the multipath load slitting. This implies the following brief set of requirements.

MP#1 It MUST be possible to identify MPLS-TP LSP.

MP#2 It MUST be possible to completely exclude MPLS-TP LSP from the multipath hash and load split.

MP#3 It SHOULD be possible to insure that an MPLS-TP LSP will not be moved to another component link as a result of a composite link load rebalancing operation.

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 MPLS LSP to determine at CSPF time whether a link or MPLS PSC LSP within the topology can support the MPLS-TP requirements of the LSP.

There is currently no signaling mechanism defined to support requirement MP#1. 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 an MPLS-TP LSP can be created with and Entropy Label Indicator (ELI) and entropy label (EL) below the MPLS-TP label [I-D.ietf-mpls-entropy-label].

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

MPLS-TP LSP can be carried as client LSP 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 [I-D.ietf-mpls-entropy-label]. This allows MPLS-TP LSP to be carried as client LSP within MPLS LSP and satisfies requirement MP#2 but requires that MPLS LSR be able to identify MPLS-TP LSP (requirement MP#1).

MPLS-TP traffic can be protected from an 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#2 is satisfied, but without a signaling extension, requirement MP#3 is not satisfied if there is a need to rebalance the load on any composite link carrying the MPLS server LSP. 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.

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 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 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. IANA Considerations

This memo includes no request to IANA.

6. 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].

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.

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.

[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.

[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.

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

Internet Engineering Task Force
Internet-Draft
Intended status: Standards Track
Expires: September 13, 2011

Quintin Zhao
Huawei Technology
Chao Zhou
Luyuan Fang
Cisco Systems
Lianyuan Li
China Mobile
Ning So
Verison Business
Raveendra Torvi
Juniper Networks
March 12, 2011

RSVP-TE Extension for Multi Topology Support
draft-zhao-mpls-rsvp-te-multi-topology-01.txt

Abstract

This document describes options to extend the existing MPLS signalling protocol RSVP for creating and maintaining Label Switching Paths (LSPs) in a Multi-Topology environments.

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 September 13, 2011.

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.

Table of Contents

1. Terminology

Terminology used in this document

MT-ID: A 12 bit value to represent Multi-Topology ID.

Default Topology: A topology that is built using the MT-ID value 0.

MT topology: A topology that is built using the corresponding MT-ID.

2. Introduction

In Multi-protocol Label Switching (MPLS) networks, a label may be assigned to represent a set of Forwarding Equivalent Classes (FEC) of packets and a mapping of the label and the FEC may be signaled along the path traversed by the packets. Therefore, the label switched paths are established to forward packets.

Resource reservation protocol (RSVP) is a network control protocol that may be used to enable applications to obtain different quality of service (QoS) for their data flows. However, RSVP is not a routing protocol. Rather, RSVP operates in conjunction with routing protocols.

Resource reservation protocol traffic engineering (RSVP-TE) is an extension to RSVP that supports resource reservations across an Internet Protocol (IP) network. Generally, RSVP-TE may be used to establish MPLS label switched paths (LSPs) with or without resource reservations, with consideration given to available bandwidth and a number of explicit hops. The LSPs may be setup using explicit routes. A variety of messages and procedures may be used by network elements to inform other network elements of the labels used for MPLS forwarding. The LSPs may be treated as a tunnel, which is tunneling below normal IP routing and filtering mechanisms.

A mechanism for Open Shortest Path First (OSPF) protocol to support multi-topologies (MT) in IP networks, wherein Type of Service (TOS) based metric fields are redefined and used to advertise different topologies is disclosed in P. Psenak, et.al., "Multi-Topology (MT) Routing in OSPF," RFC 4915, June 2007, which is incorporated herein by reference. Separate metrics may be associated for each TOS and may be advertised via protocol information exchange between network elements. The existing OSPF protocol is extended to support network topology changes with Multi-Topology Identifier (MT-ID).

A mechanism within Intermediate System to Intermediate System (IS-IS) to run a set of independent IP topologies for each network topology is disclosed in T. Przygienda, et.al., "M-ISIS: Multi Topology (MT) Routing in Intermediate System to Intermediate Systems (IS-ISs)", RFC 5120, February 2008, which is incorporated herein by reference. The existing IS-IS protocol is extended so that advertisements of adjacencies and reachable intermediate system within each topology are performed.

Therefore, there is a need to have systems and methods for supporting multi-topology in MPLS network and extending the RSVP-TE protocol as a signaling protocol in the MPLS network to establish and maintain traffic engineered LSP tunnel within each network topology or across network topologies. The LSP tunnel may need to follow a specific path or to reserve a certain amount of bandwidth to satisfy QoS requirements for the traffic flowing through the LSP tunnel within a specific network topology or across multiple network topologies.

MT based MPLS in general can be used for a variety of purposes such as service separation by assigning each service or a group of services to a topology, where the management, QoS and security of the service or the group of the services can be simplified and guaranteed, in-band management network "on top" of the original MPLS topology, maintain separate routing and MPLS forwarding domains for isolated multicast or IPv6 islands within the backbone, or force a subset of an address space to follow a different MPLS topology for the purpose of security, QoS or simplified management and/or operations.

One of the use of the MT based MPLS is where one class of data requires low latency links, for example Voice over Internet Protocol (VoIP) data. As a result such data may be sent preferably via physical landline rather than, for example, high latency links such as satellite links. As a result an additional topology is defined as all low latency links on the network and VoIP data packets are assigned to the additional topology. Another example is security-critical traffic which may be assigned to an additional topology for non-radiative links. Further possible examples are file transfer protocol (FTP) or SMTP (simple mail transfer protocol) traffic which can be assigned to additional topology comprising high latency links, Internet Protocol version 4 (IPv4) versus Internet Protocol version 6 (IPv6) traffic which may be assigned to different topology or data to be distinguished by the quality of service (QoS) assigned to it.

3. Application Scenarios

3.1. Simplified Data-plane

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

3.2. Automation of inter-layer interworking

With (G)MPLS-RSVP-MT extensions, an ingress-PE can signal particular path (ERO) that can traverses different network layer to reach a egress-PE. For instance, an ERO is associated with MT-ID RSVP subobject to indicate a "P" router to use a particular Layer-1 TE-link-state topology, instead of default Layer-3 link-state topology as illustrated in the following diagram. With this mechanism an (G)MPLS-TE LSP can be offloaded to lower layers without service disruption and without complexity of configuration.

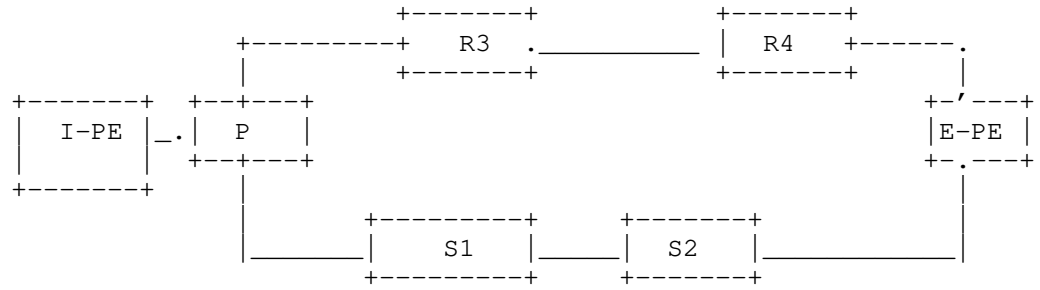


Figure 1: Layer-3 Link State Topology

Layer-3 ERO : P[MT-0]->R3->R4->E-PE[MT-0].

Inter-layer ERO : P[MT-0]->loose-hop[MT-1]->E-PE[MT-0]

Procedures to discover MT mapping with an IGP topology at ingress-PE

nodes requires some auto-discovery mechanism.

Figure 1: Layer-3 Link State Topology

3.3. Migration without service disruption

As state above, MPLS-MT abstracts link state topology and identifies it by a unique MT-ID, which need not be same as IGP-MT ID. This characteristic is quite useful for service providers looking to migrate to different flavor of IGP, e.g., OSPFv2 to ISIS6, OSPFv2 to OSPFv3. Service providers would like to incrementally upgrade the topologies, which requires an LSP to traverse multiple IGP domains (OSPFv2 to OSPFv3) or (OSPF to ISIS). In order migrate TE-LSPs to use newly deployed link state topology requires a non-trivial effort. This migration may involve service disruption, especially when a path include loose-hops in the ERO. For example: When an incoming PATH message requires an LSR to resolve loose-hop over newly deployed IGP domain, which is not possible in the absence of MPLS-MT signaling. MPLS-MT allows an ingress-PE to specify multi-topology to be used at every hop.

3.4. Service Separation

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

3.5. simplified Inter Domain TE LSP Setup

When the TE lsp is crossing multiple domains, the LSP setup process can be simplified by configuring a set of routers which are in different domains into a new single domain with a new toplogy ID using the RSVP-TE multiple topology. All the routers belong this new topology will be used to carry the traffic acrossing multiple domains and since they are in a sinle domain, so the TE lsp set up can be done easily.

3.6. Simplified inter-AS VPN Solution

When the TE lsp is crossing multiple domains for the inter-as VPN scenarios, the LSP setup process can be simplified by configuring a set of routers which are in different domains into a new single domain with a new toplogy ID using the LDP multiple topology. All

the routers belong this new topology will be used to carry the traffic acrossing multiple domains and since they are in a sinle domain with the new topology ID, so the TE lsp set up can be done easily without the complex inter-as VPN solution's option A, option B and option C.

4. Associating a RSVP message with MT-ID

RSVP-TE objects may be utilized to indicate MT information by adding the multi-topology information in an RSVP-TE object carried in a RSVP-TE message.

A preferred RSVP-TE object may be a session object.

The capability for supporting multi-topology in RSVP can be advertised during RSVP session initialization stage by including the extended RSVP session object in the first RSVP path message. After RSVP session is established, the following Path, Resv, PathErr, ResvErr and ResvConf messages will include the session object in each message and the MT ID contained in the session object will let the receiver of the message to know which topology this message is for.

This section describes an approach to associate a RSVP message with MT-ID specified in the session object.

4.1. Session Object

4.1.1. P2P LSP TUNNEL IPv4 Session Object

Class = SESSION, LSP_TUNNEL_IPv4 C-Type = 7

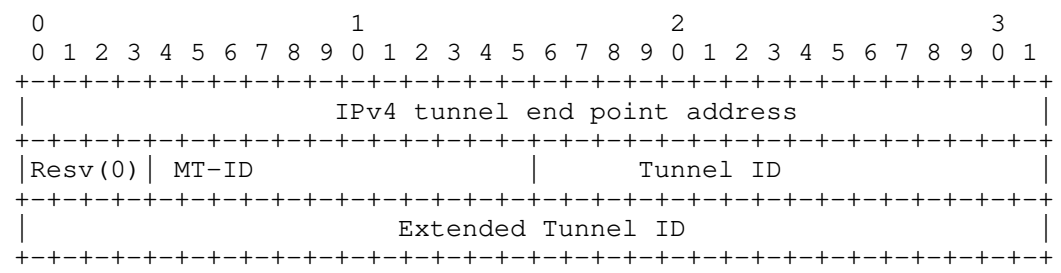


Figure 2: Format of P2P LSP_TUNNEL_IPv4 Session Object Body with MT-ID

IPv4 tunnel end point address

IPv4 address of the egress node for the tunnel.

MT-ID

A 12 bit value to represent Multi-Topology Identifier.

Tunnel ID

A 16-bit identifier used in the SESSION that remains constant over the life of the tunnel.

Extended Tunnel ID

A 32-bit identifier used in the SESSION that remains constant over the life of the tunnel. Normally set to all zeros. Ingress nodes that wish to narrow the scope of a SESSION to the ingress-egress pair may place their IPv4 address here as a globally unique identifier.

4.1.2. P2P LSP TUNNEL IPv6 Session Object

This is the same as the P2MP IPv4 LSP SESSION object with the difference that the extended tunnel ID may be set to a 16-byte identifier [RFC3209].

Class = SESSION, LSP_TUNNEL_IPv6 C_Type = 8

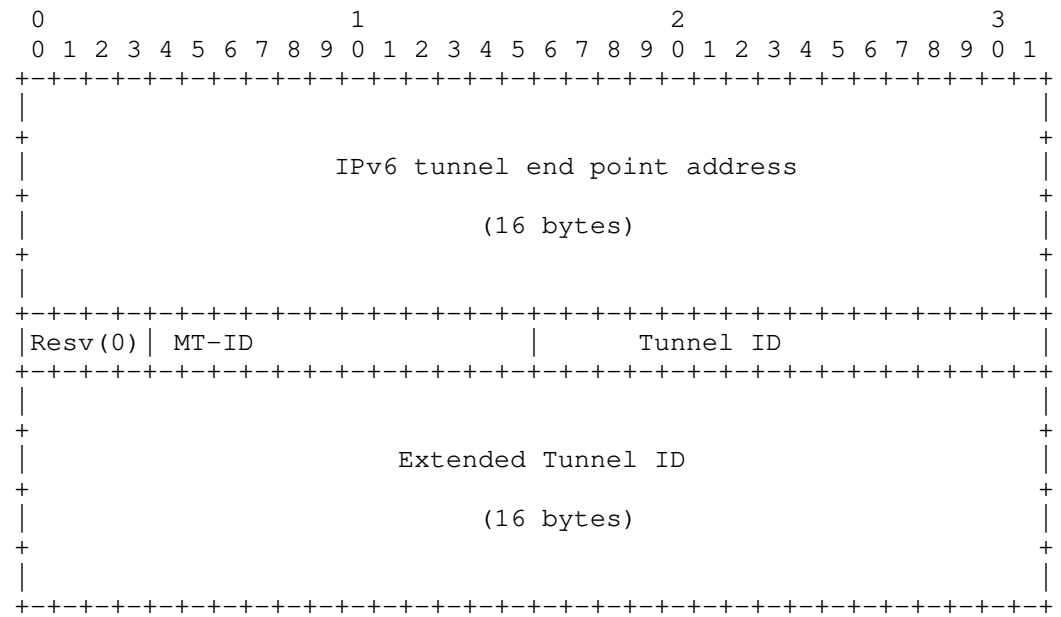


Figure 3: Format of P2P LSP_TUNNEL_IPv6 Session Object Body with MT-ID

IPv6 tunnel end point address

IPv6 address of the egress node for the tunnel.

MT-ID

A 12 bit value to represent a Multi-Topology Identifier.

Tunnel ID

A 16-bit identifier used in the SESSION that remains constant over the life of the tunnel.

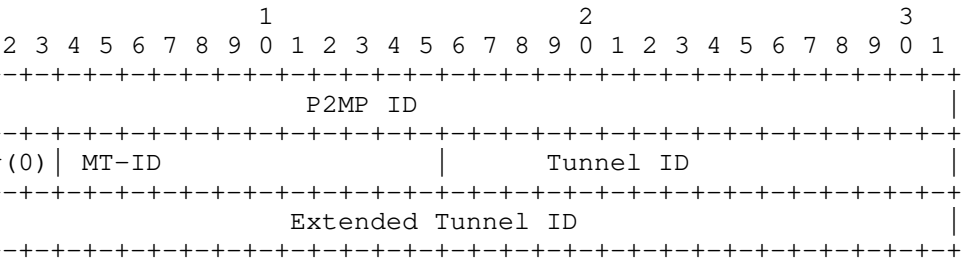
Extended Tunnel ID

A 16-byte identifier used in the SESSION that remains constant over the life of the tunnel. Normally set to all zeros. Ingress nodes that wish to narrow the scope of a SESSION to the ingress-egress pair may place their IPv6 address here as a globally unique identifier.

4.1.3. P2MP LSP TUNNEL IPv4 Session Object

This is the same as the P2MP IPv4 LSP SESSION object with the difference that the extended tunnel ID may be set to a 16-byte identifier [RFC3209].

Class = SESSION, P2MP_LSP_TUNNEL_IPv4 C-Type = 13



P2MP ID
A 32-bit identifier used in the SESSION object that remains constant over the life of the P2MP tunnel. It encodes the P2MP Identifier that is unique within the scope of the ingress LSR.

MT-ID
A 12 bit value to represent a Multi-Topology Identifier.

Tunnel ID
A 16-bit identifier used in the SESSION object that remains constant over the life of the P2MP tunnel.

Extended Tunnel ID
A 32-bit identifier used in the SESSION object that remains constant over the life of the P2MP tunnel. Ingress LSRs that wish to have a globally unique identifier for the P2MP tunnel SHOULD place their tunnel sender address here. A combination of this address, P2MP ID, and Tunnel ID provides a globally unique identifier for the P2MP tunnel.

Figure 4: Format of P2MP LSP_TUNNEL_IPv4 Session Object Body with MT-ID

4.1.4. P2MP LSP TUNNEL IPv6 Session Object

Class = SESSION, P2MP_LSP_TUNNEL_IPv6 C-Type = 14

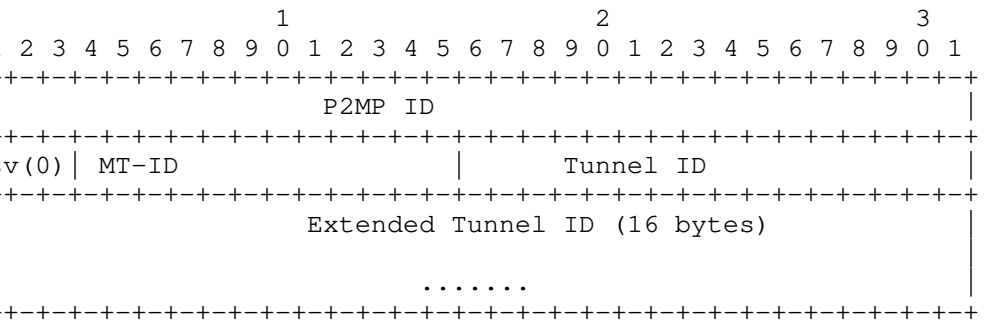


Figure 5: Format of P2MP LSP_TUNNEL_IPv6 Session Object Body with MT-ID

5. Processing of Message with MT ID

Procedure changes for processing P2P and P2MP protocol messages with MT ID: [TBD]

6. MPLS Forwarding in MT

In MT based MPLS network, forwarding will not only be based on label, but also based on the MT-ID associated with the label. There are multiple options to do this. Below, we list three options.

6.1. Use Label for (FEC, MT-ID) Tuple

The first option we propose is that MPLS forwarding for different topologies is implied by labels. This approach does not need any changes to the exiting MPLS hardware forwarding mechanism. It also resolves the forwarding issue that exists in IGP multi-topology forwarding when multiple topologies share an interface with overlaying addresses.

On a MT aware LSR, each label is associated with tuple: (FEC, MT-ID). Therefore, same FEC with different MT-ID would be assigned to different labels.

Using this option, for tuple (FEC-F, MT-ID-N1) and (FEC-F, MT-ID-N2),

each LSR along the path that is shared by topology MT-ID-N1 and MT-ID-N2 will allocate different labels to them. Thus two distinguished Label Switching Paths will be created. One (FEC-F, MT-ID-N1) and the other for (FEC-F, MT-ID-N1). The traffic for them will follow different Label Switching Paths (LSPs).

Note, in this option, label space is not allowed to be overlapping among different MTs. In the above example, each label belongs to a specific topology or the default topology. MPLS forwarding will be performed exactly same as non-MT MPLS forwarding: using label to find output information. This option will not require any change of hardware forwarding to commodate MPLS MT.

Note, We have different RIBs corresponding to different MT IDs. But we will only need one LFIB.

Below is an example for option one:

```
RIB(x) for MT-IDx:
        FEC                         NEXT HOP
        FECi(Destination A)         R1

RIB(y) for MT-IDy:
        FEC                         NEXT HOP
        FECi(Destination A)         R1

LFIB:
        Ingress Label  Egress Label       NEXT HOP
        Lm             Lp                 R1
        Ln             Lq                 R2 (could be same as R1)
```

Figure 6: FIB Entry Example for One Label Space

6.2. Overlapping Label Spaces for MT

In the option 2, label spaces are overlapping with each other, which means same label value could be used for different MT. In this option, MPLS forwarding will use label value and the MT associated with label. Each label forwarding entry will have an extra label stacked with the original label. This extra label is used as the MT identifier. For example, the forwarding entry in the LIB looks like this:

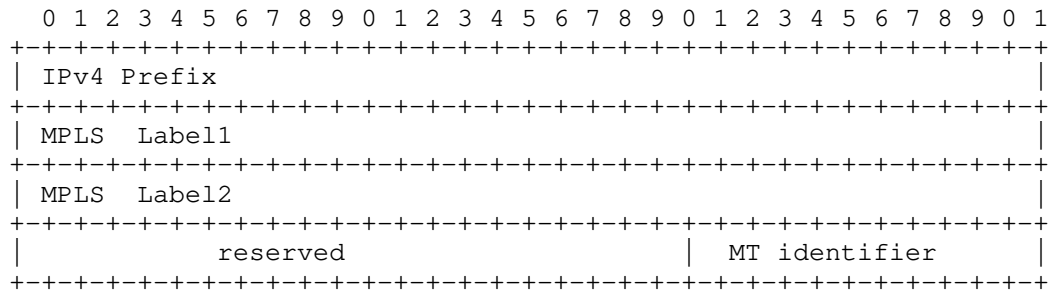


Figure 7: FIB Entry of Overlapping Label Spaces for MT

Option 1 is good for backward compatibility and it doesn't require hardware change. The disadvantage is that the 20 bits of label space is shared by all the MTs and label space for each MT is limited. The advantage for option 2 is that each MT can have full label space. The disadvantage is that they need hardware support to perform MPLS MT forwarding. In addition, option 2 require one more label lookup.

7. Reserved MT ID Values

Certain MT topologies are assigned to serve pre-determined purposes: [TBD]

8. Security Consideration

MPLS security applies to the work presented. No specific security issues with the proposed solutions are known. The authentication procedure for RSVP signalling is the same regardless of MT information inside the RSVP messages.

9. IANA Considerations

TBD

10. Acknowledgement

The authors would like to thank Dan Tappan, Nabil Bitar and Huang Xin for their valuable comments on this draft.

11. References

11.1. Normative References

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

[RFC3692] Narten, T., "Assigning Experimental and Testing Numbers Considered Useful", BCP 82, RFC 3692, January 2004.

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

[RFC5120] Przygienda, T., Shen, N., and N. Sheth, "M-ISIS: Multi Topology (MT) Routing in Intermediate System to Intermediate Systems (IS-ISs)", RFC 5120, February 2008.

[RFC3209] Awduche, D., Berger, L., Gan, D., Li, T., Srinivasan, V., and G. Swallow, "RSVP-TE: Extensions to RSVP for LSP Tunnels", RFC 3209, December 2001.

[RFC4420] Farrel, A., Papadimitriou, D., Vasseur, J., and A. Ayyangar, "Encoding of Attributes for Multiprotocol Label Switching (MPLS) Label Switched Path (LSP) Establishment Using Resource ReserVation Protocol-Traffic Engineering (RSVP-TE)", RFC 4420, February 2006.

[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.

11.2. Informative References

Authors' Addresses

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

Email: qzhao@huawei.com

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

Email: huaimochen@huawei.com

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

Email: Ning.So@verizonbusiness.com

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

Email: lufang@cisco.com

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

Email: czhou@cisco.com

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

Email: lilianyuan@chinamobile.com

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

Email: huanglu@chinamobile.com

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

Email: lichenyj@chinamobile.com

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

Email: pratiravi@juniper.com

Network working group
Internet Draft
Intended status: Standards Track
Updates: RFC 5036 (if approved)
Expires: September 2011

L. Zheng
M. Chen
Huawei Technologies

March 14, 2011

LDP Hello Cryptographic Authentication

draft-zheng-mpls-ldp-hello-crypto-auth-01.txt

Status of this Memo

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

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

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

The list of current Internet-Drafts can be accessed at http://www.ietf.org/ietf/1id-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 September 14, 2011.

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

This document introduces a new Cryptographic Authentication TLV which is used in LDP Hello message as an optional parameter. It enhances the authentication mechanism for LDP by securing the Hello message against spoofing attack.

Conventions used in this document

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

Table of Contents

1. Introduction

The Label Distribution Protocol (LDP) [RFC 5036] utilizes LDP sessions that run between LDP peers. The peers may be directly connected at the link level or may be remote. A label switching router (LSR) that speaks LDP may be configured with the identity of its peers or may discover them using the LDP Hello message sent encapsulated in UDP that may be addressed to "all routers on this subnet" or to a specific IP address. Periodic Hello messages are

also used to maintain the relationship between LDP peers necessary to keep the LDP session active.

Unlike all other LDP messages, the Hello messages are sent using UDP not TCP. This means that they cannot benefit from the security mechanisms available with TCP. [RFC5036] does not provide any security mechanisms for use with Hello messages except to note that some configuration may help protect against bogus discovery events.

Spoofing a Hello packet for an existing adjacency can cause the valid adjacency to time out and in turn can result in termination of the associated session. This can occur when the spoofed Hello specifies a smaller Hold Time, causing the receiver to expect Hellos within this smaller interval, while the true neighbor continues sending Hellos at the previously agreed lower frequency. Spoofing a Hello packet can also cause the LDP session to be terminated directly, which can occur when the spoofed Hello specifies a different Transport Address, other than the previously agreed one between neighbors. Spoofed Hello messages is observed and reported as real problem in production networks.

As described in [RFC5036], the threat of spoofed Basic Hellos can be reduced by accepting Basic Hellos only on interfaces to which LSRs that can be trusted, and ignoring Basic Hellos not addressed to the "all routers on this subnet" multicast group. Spoofing attacks via Extended Hellos are potentially more serious threat. An LSR can reduce the threat of spoofed Extended Hellos by filtering them and accepting only those originating at sources permitted by an access list. However, performing the filtering using access lists requires LSR resource, and the LSR is still vulnerable to the IP source address spoofing.

This document introduces a new Cryptographic Authentication TLV which is used in LDP Hello message as an optional parameter. It enhances the authentication mechanism for LDP by securing the Hello message against spoofing attack, and an LSR can be configured to only accept Hello messages from specific peers when authentication is in use.

Using this Cryptographic Authentication TLV, one or more secret keys (with corresponding key IDs) are configured in each system. For each LDP Hello packet, the key is used to generate and verify a HMAC Hash that is stored in the LDP Hello packet. For cryptographic hash function, this document proposes to use SHA-1, SHA-256, SHA-384, and SHA-512 defined in US NIST Secure Hash Standard (SHS) [FIPS-180-3]. The HMAC authentication mode defined in NIST FIPS 198 is used [FIPS-

198]. Of the above, implementations MUST include support for at least HMAC-SHA-256 and SHOULD include support for HMAC-SHA-1 and MAY include support for either of HMAC-SHA-384 or HMAC-SHA-512.

2. Cryptographic Authentication TLV

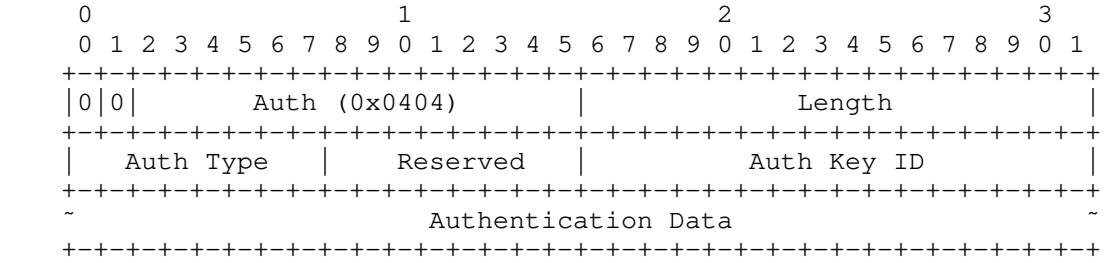
2.1. Optional Parameter for Hello Message

[RFC5036] defines the encoding for the Hello message. Each Hello message contains zero or more Optional Parameters, each encoded as a TLV. Three Optional Parameters are defined by [RFC5036]. This document defines a new Optional Parameter: the Cryptographic Authentication parameter.

Optional Parameter	Type
IPv4 Transport Address	0x0401 (RFC5036)
Configuration Sequence Number	0x0402 (RFC5036)
IPv6 Transport Address	0x0403 (RFC5036)
Cryptographic Authentication	0x0404 (this document, TBD by IANA)

The Cryptographic Authentication TLV Encoding is described in section 2.2.

2.2. Cryptographic Authentication TLV Encoding



- Type: 0x0404 (TBD by IANA), Cryptographic Authentication

- Length: Specifying the length in octets of the value field.

- Auth Type: The authentication type in use

0 - HMAC-SHA-1
1 - HMAC-SHA-256
2 - HMAC-SHA-384
3 - HMAC-SHA-512
4-255 - Reserved for future use
(TBD by IANA)

- Reserved: MUST be set to zero on transmit, and ignored on receipt

- Auth Key ID: The authentication key ID in use for this packet. This allows one or more keys to be active simultaneously.

- Authentication Data:

  This field carries the digest computed by the Cryptographic Authentication algorithm in use. The length of the Authentication Data varies based on the cryptographic algorithm in used, which is shown as below:

Auth type	Length
HMAC-SHA1	20 bytes
HMAC-SHA-256	32 bytes
HMAC-SHA-384	48 bytes
HMAC-SHA-512	64 bytes

3. Cryptographic Aspects

In the algorithm description below, the following nomenclature, which is consistent with [FIPS-198], is used:

- H is the specific hashing algorithm specified by Auth Type (e.g. SHA-256).

- K is the Authentication Key for the Hello packet.

- Ko is the cryptographic key used with the hash algorithm.

- B is the block size of H, in octets.

  For SHA-1 and SHA-256: B == 64

  For SHA-384 and SHA-512: B == 128

- L is the length of the hash outputs, in octets.

- XOR is the exclusive-or operation.

- Ipad is the byte 0x36 repeated B times.

- Opad is the byte 0x5c repeated B times.

- Apad is the byte 0x878FE1F3 repeated (L/4) times.

3.1. Cryptographic Key

As described in RFC 2104, the authentication key K can be of any length up to B. Applications that use keys longer than B bytes will first hash the key using H and then use the resultant L byte string as the actual key to HMAC.

In this application, Ko is always L octets long. If the Authentication Key (K) is L octets long, then Ko is equal to K. If the Authentication Key (K) is more than L octets long, then Ko is set to H(K). If the Authentication Key (K) is less than L octets long, then Ko is set to the Authentication Key (K) with trailing zeros such that Ko is L octets long.

3.2. Hash

First, the Authentication Data field in the Cryptographic Authentication TLV is filled with the value Apad and the Auth Type field is set accordingly per Cryptographic Authentication algorithm in use.

Then, to compute HMAC over the Hello packet it performs:

```
H(Ko XOR Opad || H(Ko XOR Ipad || (Hello Packet)))
```

Hello Packet here is the entire LDP Hello packet including the IP header.

3.3. Result

The resultant Hash becomes the Authentication Data that is sent in the Authentication Data field of the Cryptographic Authentication TLV. The length of the Authentication Data field is always identical to the message digest size of the specific hash function H that is being used.

4. Processing Hello Message Using Cryptographic Authentication

4.1. Transmission Using Cryptographic Authentication

Prior to transmitting Hello message, the Auth Type field is set to indicate the authentication type in use. The Length in the Cryptographic Authentication TLV header is set as per the authentication algorithm that is being used. It is set to 24 for HMAC-SHA-1, 36 for HMAC-SHA-256, 52 for HMAC-SHA-384 and 68 for HMAC-SHA-512.

The Auth Key ID field is set to the ID of the current authentication key. The HMAC Hash is computed as explained in Section 3. The resulting Hash is stored in the Authentication Data field prior to transmission. The authentication key MUST NOT be carried in the packet.

4.2. Receipt Using Cryptographic Authentication

The receiving LSR applies acceptability criteria for received Hellos using cryptographic authentication. If the Cryptographic Authentication TLV is unknown to the receiving LSR, the received packet MUST be discarded according to Section 3.5.1.2.2 of [RFC5036].

If the Cryptographic Authentication TLV in a received Hello packet does not contain a known and acceptable Auth Type value, then the received packet MUST be discarded. If the Auth Key ID field does not match the ID of a configured authentication key, the received packet MUST be discarded.

Before the receiving LSR performs any processing, it needs to save the values of the Authentication Data field. The receiving LSR then replaces the contents of the Authentication Data field with Apad, computes the Hash, using the authentication key specified by the

received Auth Key ID field, as explained in Section 3. If the locally computed Hash is equal to the received value of the Authentication Data field, the received packet is accepted for other normal checks and processing as described in [RFC5036]. Otherwise, the received packet MUST be discarded.

5. Security Considerations

Section 1 of this document describes the security issues arising from the use of unsecured LDP Hello messages. In order to combat those issues, it is RECOMMENDED that all deployments use the Cryptographic Authentication TLV to secure the Hello message.

The quality of the security provided by the Cryptographic Authentication TLV depends completely on the strength of the cryptographic algorithm in use, the strength of the key being used, and the correct implementation of the security mechanism in communicating LDP implementations. Also, the level of security provided by the Cryptographic Authentication TLV varies based on the authentication type used.

6. IANA Considerations

IANA maintains a registry of LDP message parameters with a sub-registry to track LDP TLV Types. This document request IANA to assign a new TLV Types as follows:

TLV	Type
Cryptographic Authentication	0x0404 (TBD)

This document also request IANA to assign a new registry titled "LDP Hello Authentication Type", its recommended values as follows:

Value	LDP Hello Authentication Type Name
0	HMAC-SHA1
1	HMAC-SHA-256
2	HMAC-SHA-384
3	HMAC-SHA-512
4-255 (TBD)	Unassigned

7. Acknowledgments

The authors would like to thank Liu Xuehu for his work on background and motivation for LDP Hello authentication. The authors also would like to thank Adrian Farrel, Thomas Nadeau, So Ning, Eric Rosen, Sam Hartman and Manav Bhatia for their valuable comments.

8. References

8.1. Normative References

[RFC2104] Krawczyk, H. et al., "HMAC: Keyed-Hashing for Message Authentication", RFC 2104, February 1997.

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

[RFC5036] Andersson, L., Minei, I., and B. Thomas, "LDP Specification", RFC 5036, October 2007.

[FIPS-180-3] National Institute of Standards and Technology, "Secure Hash Standard (SHS)", FIPS PUB 180-3, October 2008.

[FIPS-198] US National Institute of Standards & Technology, "The Keyed-Hash Message Authentication Code (HMAC)", FIPS PUB 198, March 2002.

8.2. Informative References

[RFC2385] Heffernan, A., "Protection of BGP Sessions via the TCP MD5 Signature Option", RFC 2385, August 1998.

[RFC4634] Eastlake 3rd, D. and T. Hansen, "US Secure Hash Algorithms (SHA and HMAC-SHA)", RFC 4634, July 2006.

[RFC5709] Bhatia, M., Manral, V., Fanto, M., White, R., Barnes, M., Li, T., and R. Atkinson, "OSPFv2 HMAC-SHA Cryptographic Authentication", RFC 5709, October 2009.

[RFC5880] Katz, D. and D. Ward, "Bidirectional Forwarding Detection", RFC 5880, June 2010.

Authors' Addresses

Lianshu Zheng
Huawei Technologies Co., Ltd.
Huawei Building, No.3 Xinxi Road,
Hai-Dian District,
Beijing 100085
China

Email: verozheng@huawei.com

Mach(Guoyi) Chen
Huawei Technologies Co., Ltd.
Huawei Building, No.3 Xinxi Road,
Hai-Dian District,
Beijing 100085
China

Email: mach@huawei.com