

MPLS Working Group
Internet Draft
Intended status: Standard Track
Expires: January 6, 2012

Zafar Ali
Rakesh Gandhi
Cisco Systems, Inc.
July 7, 2011

Signaling RSVP-TE P2MP LSPs in an Inter-domain Environment
draft-ali-mps-inter-domain-p2mp-rsvp-te-lsp-06.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). 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 6, 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.

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

Point-to-MultiPoint (P2MP) Multiprotocol Label Switching (MPLS) and Generalized MPLS (GMPLS) Traffic Engineering Label Switched Paths (TE LSPs) may be established using signaling techniques described in [RFC4875]. However, [RFC4875] does not address many issues that comes when a P2MP-TE LSP is signaled in an inter-domain networks. Specifically, one of the issues in inter-domain networks is how to allow computation of a loosely routed P2MP-TE LSP such that it is re-merge free. This document provides a framework and required protocol extensions needed for establishing and controlling P2MP MPLS and GMPLS TE LSPs in inter-domain networks.

This document borrows inter-domain TE terminology from [RFC 4726], e.g., for the purposes of this document, a domain is considered to be any collection of network elements within a common sphere of address management or path computational responsibility. Examples of such domains include Interior Gateway Protocol (IGP) areas and Autonomous Systems (ASes).

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

Copyright Notice.....1
1. Introduction.....3
2. Framework.....5

2.1. Signaling Options.....	5
2.2. Path Computation Techniques.....	5
3. Control Plane Solution.....	5
3.1. Single Border Node.....	6
3.2. Crankback and Path Error Signaling Procedure.....	6
4. Data Plane Solution.....	7
4.1. P2MP-TE Re-merge Recording Request Flag.....	7
4.2. P2MP-TE Re-merge Present Flag.....	8
4.3. Signaling Procedure.....	9
5. Security Considerations.....	10
6. IANA Considerations.....	10
7. Acknowledgments.....	11
8. References.....	11
8.1. Normative References.....	11
8.2. Informative References.....	11
Author's Addresses.....	12

1. Introduction

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

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

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

This requirement becomes more acute to address when P2MP LSP spans multiple domains. For the purposes of this document, a domain is considered to be any collection of network elements

within a common sphere of address management or path computational responsibility. Examples of such domains include Interior Gateway Protocol (IGP) areas and Autonomous Systems (ASes). This is because in an inter-domain environment, the ingress node may not have topological visibility into other domains to be able to compute and signal a re-merge free P2MP LSP. In that case, the border node for a new domain will be given one or more loose next hops for the P2MP LSP. When processing a path message, it may not have knowledge of all of the destinations of the P2MP LSP, either because S2L sub-LSPs are split between multiple Path messages, or because not all S2L sub-LSPs pass through this border node. In that case, existing protocol mechanisms do not provide sufficient information for it to be able to expand the loose hop(s) in such a way that the overall P2MP path is guaranteed to be optimal and re-merge free.

RFC 4875 specifies two approaches to handle re-merge conditions. In the first method that is based on control plane handling, the re-merge node initiates the removal of the re-merge branch(es) by sending a Path Error message. In the second method that is based on data plane handling, the node detecting the re-merge case, i.e., the re-merge node, allows the re-merge to persist, but data from all but one incoming interface is dropped at the re-merge node. This ensures that duplicate data is not sent on any outgoing interface.

This document proposes RSVP-TE signaling procedures for P2MP LSP to handle re-merge for both using control plane approach and data plane approach.

Control plane solution is using crankback signaling in RSVP. [RFC5151] describes mechanisms for applying crankback to inter-domain P2P LSPs, but does not cover P2MP LSPs. Also, crankback mechanisms for P2MP LSPs are not addressed by [RFC4875]. This document describes how crankback signaling extensions for MPLS and GMPLS RSVP-TE defined in [RFC4920] can be used for setting up P2MP TE LSPs to resolve re-merges.

Data plane solution described in [RFC4875] is extended by using a new flag in RRO Attributes Sub-object in RSVP. The proposed solution makes use of RRO Attributes Sub-object as defined in [RFC5420] for this purpose. This document describes how new RRO Attributes Flag can be used to handle P2MP re-merge conditions efficiently.

The solutions presented in this document do not guarantee optimization of the overall P2MP tree across all domains. PCE can

be used, instead, to address optimization of the overall P2MP tree.

2. Framework

2.1. Signaling Options

The four signaling options defined for P2P inter-domain LSPs in [RFC4726] are also applicable to P2MP LSPs.

- . LSP nesting, using hierarchical LSPs [RFC4206].
- . A single contiguous LSP, using the same SESSION and LSP ID along its whole path.
- . LSP stitching [RFC5150].
- . A combination of the above.

In the case of LSP nesting using hierarchical LSPs, the tunneled LSP MUST use upstream-assigned labels to ensure that the same label is used at every leaf of the H-LSP ([RFC5331], [I-D.ietf-mppls-rsvp-upstream]). The H-LSP SHOULD request non-PHP behavior and out-of-band mapping as defined in [I-D.ietf-mppls-rsvp-te-no-php-oob-mapping].

2.2. Path Computation Techniques

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

3. Control Plane Solution

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

3.1. Single Border Node

The ingress node is RECOMMENDED to select the same border node as an ERO loose hop for all sibling S2L sub-LSPs that transit a given domain. This reduces the chances of the sibling S2L sub-LSPs in remerging states, because a single border node has the necessary state to ensure that the path that they take through the domain is re-merge free.

3.2. Crankback and Path Error Signaling Procedure

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

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

If a node on the path of the P2MP LSP is unable to find a route that can supply the required resources or that is re-merge free, it SHOULD generate a Path Error message for the subset of the S2L sub-LSPs which it is not able to route. For this purpose the node SHOULD try to find a minimum subset of S2L sub-LSPs for which the Path Error needs to be generated. This rule applies equally to the case where multiple S2L sub-LSPs are signaled using one Path message, as to the case where a single S2L sub-LSP is signaled in each Path message. RSVP-TE Notify messages do not include S2L_SUB_LSP objects and cannot be used to send errors for a subset of the S2L sub-LSPs in a Path message. For that reason, the node SHOULD use a Path Error message rather than a Notify message to communicate the error. In the case of a re-merge error, the node SHOULD use the error code "Routing Problem" and the error value "ERO resulted in re-merge" as specified in [RFC4875].

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

[RFC5151]. If all subsequent attempts by the border node are unsuccessful, the border node MUST send the held Path Error upstream to the ingress node.

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

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

As discussed in [RFC4090] section 3.3, border node MAY keep the history of Path Errors. In case of P2MP LSPs, ingress node and border nodes may keep re-merge Path Errors in history table until S2L sub-LSPs have been successfully established or until local timer expires.

4. Data Plane Solution

As mentioned in [RFC4875], node may accept the remerging S2Ls but only send the data from one of these interfaces to its outgoing interfaces. That is, the node MUST drop data from all but one incoming interface. This ensures that duplicate data is not sent on any outgoing interface.

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

RSVP-TE signaling extensions are defined in the following to request P2MP-TE Re-merge Recording and indicate P2MP-TE Re-merge Presence.

4.1. P2MP-TE Re-merge Recording Request Flag

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

Expires January 2012

[Page 7]

^L

Internet-Draft draft-ali-mpls-inter-domain-p2mp-rsvp-te-lsp-06.txt

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

The P2MP-TE Re-merge Recording Request flag is meaningful on a Path message and can be inserted by the ingress node or a border

node.

If the P2MP-TE Re-merge Recording Flag is set to 1, it means that "P2MP-TE Re-merge Presence" defined in the next section should be used to indicate to the ingress and border nodes along the setup of the LSP that a remerge is present but accepted and that incoming traffic is being dropped for the given S2L.

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

4.2. P2MP-TE Re-merge Present Flag

The P2MP-TE Re-merge Present Flag is the counter part of the P2MP-TE Re-merge Recording Request flag defined above. Specifically, RSVP signaling extension is defined to indicate to the upstream node of the re-merge condition and that incoming traffic is being dropped for the given S2L.

When a node decides to accept remerge and drop traffic from an incoming interface for an S2L due to the re-merge condition, and understands the "P2MP-TE Re-merge Recording Request in the Attribute Flags TLV of the LSP_ATTRIBUTES object of the Path message, the node SHOULD set the newly defined "P2MP-TE Re-merge Present" flag in the RRO Attributes sub-object defined in [RFC 5420] in RRO.

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

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

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

4.3. Signaling Procedure

When a node receives an S2L sub-LSP Path message with LSP Attributes sub-object that has "P2MP-TE Re-merge Recording Request" Flag set, and the node does not support data plane based re-merge handling, and the S2L is causing a re-merge, the node SHOULD reject the S2L sub-LSP path message and send the Path

Error with the error code "Routing Problem" and the error value "ERO resulted in re-merge" as specified in [RFC4875].

When a path message is received at a transit node and "P2MP-TE Re-merge Recording Request" Flag is set in the LSP Attributes sub-object, the node MAY decide to accept the re-merge S2L sub-LSP. In this case, before the Resv message is sent to the upstream node, the node adds the RRO Attributes sub-object to the RRO and sets the "P2MP-TE Re-merge Recording Request" Flag. .

When a transit node receives a reservation message for an S2L that is causing a re-merge, the node SHOULD set the "P2MP-TE Re-merge Present" flag in the RRO Attributes sub-object in the reservation message if it decides to drop the incoming traffic of this S2L. "P2MP-TE Re-merge Present" flag in RRO Attribute sub-object is not set for the S2Ls if the node has selected the incoming interface of the S2Ls to forward the traffic.

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

Proposed signaling extensions allow an ingress node and a border node to have a complete view of the re-merges on entire S2L path and on all S2Ls of the P2MP tree and can take appropriate actions to resolve re-merges and optimize network bandwidth resources. The proposed signaling extensions are equally applicable to single domain scenarios.

A node may need to select a different incoming interface to forward traffic in future. In that case, a reservation change message is sent upstream indicating the change by marking or clearing the "P2MP-TE Re-merge Present" flag appropriately for all effected S2Ls.

The re-merge node SHOULD NOT dynamically change incoming interface to forward traffic to avoid unnecessary race conditions.

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

5. Security Considerations

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

6. IANA Considerations

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

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

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

- o P2MP-TE Re-merge Recording Present Flag:
 - Bit Number: To be assigned by IANA.
 - Attribute flag carried in Path message: No
 - Attribute flag carried in RRO Attributes sub-object in RRO of the Resv message: Yes

Expires January 2012

[Page 10]

^L

Internet-Draft draft-ali-mppls-inter-domain-p2mp-rsvp-te-lsp-06.txt

7. Acknowledgments

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

8. References

8.1. Normative References

- [RFC4875] Aggarwal, R., Papadimitriou, D., and S. Yasukawa, "Extensions to Resource Reservation Protocol Traffic

Engineering (RSVP-TE) for Point-to-Multipoint TE Label Switched Paths (LSPs)", RFC 4875, May 2007.

- [RFC5151] Farrel, A., Ayyangar, A., and JP. Vasseur, "Inter-Domain MPLS and GMPLS Traffic Engineering -- Resource Reservation Protocol-Traffic Engineering (RSVP-TE) Extensions", RFC 5151, February 2008.
- [RFC4920] Farrel, A., Satyanarayana, A., Iwata, A., Fujita, N., and G. Ash, "Crankback Signaling Extensions for MPLS and GMPLS RSVP-TE", RFC 4920, July 2007.
- [RFC5920] L. Fang, Ed., "Security Framework for MPLS and GMPLS Networks", RFC 5920, July 2010.
- [RFC3209] Awduche, D., Berger, L., Gan, D., Li, T., Srinivasan, V., and G. Swallow, "RSVP-TE: Extensions to RSVP for LSP Tunnels", RFC 3209, December 2001.

8.2. Informative References

- [RFC4726] Farrel, A., Vasseur, J., and A. Ayyangar, "A Framework for Inter-Domain Multiprotocol Label Switching Traffic Engineering", RFC 4726, November 2006.
- [RFC4206] Kompella, K. and Y. Rekhter, "Label Switched Paths (LSP) Hierarchy with Generalized Multi-Protocol Label Switching (GMPLS) Traffic Engineering (TE)", RFC 4206, October 2005.
- [RFC5150] Ayyangar, A., Kompella, K., Vasseur, JP., and A. Farrel, "Label Switched Path Stitching with Generalized Multiprotocol Label Switching Traffic Engineering (GMPLS TE)", RFC 5150, February 2008.

Expires January 2012

[Page 11]

^L

Internet-Draft draft-ali-mpls-inter-domain-p2mp-rsvp-te-lsp-06.txt

- [RFC5331] Aggarwal, R., Rekhter, Y., and E. Rosen, "MPLS Upstream Label Assignment and Context-Specific Label Space", RFC 5331, August 2008.
- [I-D.ietf-mpls-rsvp-upstream] Aggarwal, R. and J. Roux, "MPLS Upstream Label Assignment for RSVP-TE", draft-ietf-mpls-rsvp-upstream-05 (work in progress), March 2010.
- [I-D.ietf-mpls-rsvp-te-no-php-oob-mapping] Ali, Z. and G. Swallow, "Non PHP Behavior and out-of-band mapping for RSVP-TE LSPs", draft-ietf-mpls-rsvp-te-no-php-oob-mapping-04 (work in progress), March 2010.

Author's Addresses

Zafar Ali
Cisco Systems, Inc.
Email: zali@cisco.com

Rakesh Gandhi
Cisco Systems, Inc.
Email: rgandhi@cisco.com

Network Working Group
Internet-Draft
Intended status: Informational
Expires: January 5, 2012

T. Beckhaus
Deutsche Telekom AG
B. Decraene
France Telecom
K. Tiruveedhula
M. Konstantynowicz
Juniper Networks
July 4, 2011

LDP Downstream-on-Demand in Seamless MPLS
draft-beckhaus-ldp-dod-00

Abstract

Seamless MPLS design enables a single IP/MPLS network to scale over core, metro and access parts of a large network infrastructure using standardized IP/MPLS protocols. One of the key goals of Seamless MPLS is to meet requirements specific to access devices, based on their position in the network topology and their compute and memory constraints limit the amount of state they can hold. This can be achieved with LDP Downstream-on-Demand (LDP DoD) as specified in RFC 5036 [RFC5036]. This document describes LDP DoD use cases and lists LDP DoD procedures in the context of Seamless MPLS design.

Requirements Language

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

Status of this Memo

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

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

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

This Internet-Draft will expire on January 5, 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	3
2.	Reference Topologies	4
3.	LDP DoD Use Cases	6
3.1.	Access Node Start-Up	7
3.2.	Access Node Service Provisioning	8
3.3.	Access Node Service Decommissioning	9
3.4.	Service Failure	10
3.5.	Network Transport Failure	11
3.5.1.	Access Node Failure	11
3.5.2.	Access Node Uplink Failure	11
3.5.3.	AGN node failure	12
3.5.4.	AGN network-side reachability failure	12
4.	LDP Downstream on Demand Procedures	12
4.1.	LDP Label Distribution Modes	13
4.2.	IPv6 Support	14
4.3.	LDP DoD Session Negotiation	14
4.4.	Label Request Procedures	15
4.4.1.	AN Label Request Procedure	15
4.4.2.	AGN Label Request Procedure	16
4.4.3.	Label Request Retry Procedure	16
4.5.	Label Withdraw Procedure	17
4.6.	Label Release Procedure	18
4.7.	Local Repair	18
5.	IANA Considerations	19
6.	Security Considerations	19
7.	Acknowledgements	19
8.	References	19
8.1.	Normative References	19
8.2.	Informative References	19
	Authors' Addresses	20

1. Introduction

Seamless MPLS design enables a single IP/MPLS network to scale over core, metro and access parts of a large network infrastructure using standardized IP/MPLS protocols. One of the key goals of Seamless MPLS is to meet requirements specific to access devices, based on their position in the network topology and their compute and memory constraints limit the amount of state they can hold. This can be achieved with LDP Downstream-on-Demand (LDP DoD) as specified in RFC 5036 [RFC5036]. This document describes LDP DoD use cases and lists LDP DoD procedures in the context of Seamless MPLS design.

In Seamless MPLS topologies described in [seamless-mpls] , IP/MPLS protocol optimization is possible due to a relatively simple network topology that access nodes (AN) are part of.

AN connectivity options include:

- o AN single-homed to an aggregation node (AGN)
 - * with single link
 - * with multiple parallel links (IP ECMP or L2 LAG)
- o AN dual-homed to two AGNs
 - * with single link
 - * with multiple parallel links (IP ECMP or L2 LAG)
- o AN daisy-chained via hub-AN
 - * with single link
 - * with multiple parallel links (IP ECMP or L2 LAG)

With such topologies AN can implement the simplest IP routing configuration with static routes, limiting number of IP RIB and FIB entries required on AN. Furthermore MPLS label assignment can be addressed with LDP Downstream-on-Demand (LDP DoD) distribution. In general MPLS routers implement LDP Downstream Unsolicited (LDP DU) label distribution, advertising MPLS labels for all routes in their RIB. This is seen as very insufficient as ANs only require a small subset of total routes (and associated labels). LDP DoD enables on-request label distribution ensuring that only required labels are requested, provided and installed. Note that LDP DoD implementation is not widely available in today's IP/MPLS devices despite the fact that it has been described in the original LDP specification RFC 5036

[RFC5036]. This is because the original LDP DoD specification has been mainly used for ATM and FR-based MPLS LSRs in order to conserve available label space (i.e. with labels encoded in VPI/VCI).

2. Reference Topologies

Following reference end to end network topology is used for review the LDP DoD use cases based on Seamless MPLS [I-D.ietf-mpls-seamless-mpls]:

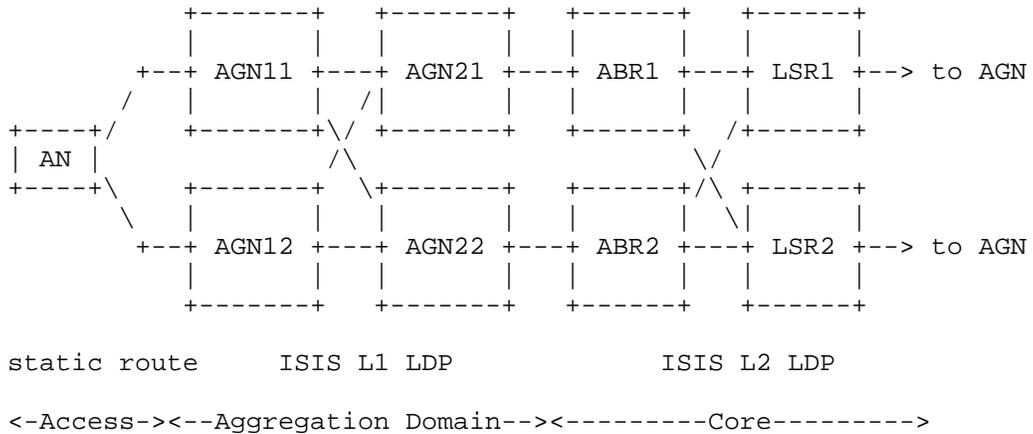


Figure 1. End-to-end reference network topology.

Access Node is either single or dual homed to AGN1(s), with either a single or multiple parallel links to AGN1(s). AN has an LDP DoD session configured between its loopback address and the loopback address(es) of AGN1s. The reference AN configuration is shown in figure below.

- o Static default routes (0.0.0.0/0) pointing to all interfaces linking to AGN1 nodes, one per interface.
- o Specific static routes for /32 loopback address of the directly connected AGN1, pointing to all interfaces linked to this AGN1.

If AN does not support RFC 5283 [RFC5283] and LDP label acceptance based on the longest match in the RIB, specific static routes to all required /32 destinations will be configured on this AN. This configuration is service dependent: every unique destination requires a distinct /32 static routing entry pointing to interfaces linking to AGN1 nodes.

AGN1s have specific /32 static routes for adjacent AN's loopback address, pointing to all interfaces linked to this AN. If interfaces to AN are up, AGN1 advertises this /32 FEC over LDP Downstream Unsolicited (LDP DU) sessions to AGN2s.

Note: Additional LDP features should be supported to comply with Seamless MPLS fast service restoration requirements as follows:

- o AN should support local-repair in case of AGN1 uplink or AGN1 failure, by using either LDP LFA or simple ECMP or primary/backup switchover.
- o AGN1 should support local-repair in case of AN downlink failure, by implementing IGP LFA (w/ LDP support).
- o AGN should be configured with LDP-IGP synchronization to avoid traffic loss where there is no LDP label allocated to the downstream best IGP next hop.
- o AN and AGN should be configured with LDP session protection to avoid delay upon the recovery from link failure.

3. LDP DoD Use Cases

LDP DoD operation is driven by Seamless MPLS use cases. This section describes these use cases focusing on services provisioned on Access Node and required LDP DoD operation on AN and AGN. For simplicity an example of MPLS PWE3 service is used to illustrate the service use cases.

Described LDP DoD operations apply equally to ANs connected over single links or parallel links (IP ECMP or L2 LAG).

This document is focusing on IPv4 LDP DoD procedures. Similar

procedures are required with IPv6 LDP DoD, however some extension specific to IPv6 will apply including LSP mapping, peer discovery, transport connection establishment. These will be added in this document once LDP IPv6 standardization is advanced as per [I-D.ietf-mpls-ldp-ipv6].

3.1. Access Node Start-Up

Access Node (AN) is commissioned without any service provisioned. AN may request labels for loopback addresses of AGN1, AGN2 or other nodes within Seamless MPLS. During the initial AN configuration no services are provisioned on it. It is assumed that AGN1 has required IP/MPLS configuration for network-side connectivity.

AN is only provisioned with the following static IP routing entries:

- o Static default route 0/0 pointing to interfaces connected to AGN1 (AGN11 and AGN12 if present).
- o Static route for adjacent AGN1's loopback, pointing to interfaces connected to AGN1.
- o (Optional) Static route for local metro AGN2's loopback, pointing to interfaces connected to AGN1.
- o (Optional) Static routes for other nodes within Seamless MPLS network.

Note: last two entries are optional and not required if AN supports inter-area LDP RFC 5283 [RFC5283] and triggering of LDP DoD label mapping request by service (e.g. PW) configuration.

IP/MPLS configuration on AN includes LDP sessions to loopback addresses of adjacent AGN1's. Source IP address for this LDP session is the loopback address of AN.

AGN1 is provisioned with a static route for AN's loopback, pointing to the interface connected to this AN. When the interface and link are up, AGN1 advertises this AN's static route into network side routing protocols i.e. IGP and/or MP-BGP Labeled Unicast.

Access Node SHOULD request labels over LDP DoD session(s) to AGN1(s) for all configured static routes if this static routes are configured with LDP DoD request policy. It is expected that all /32 static routes will be configured with such policy.

AGN1 provides AN with requested labels and MUST install the labels in its label table (LIB) and its forwarding table (LFIB). Access Node

MUST also install the labels in its LIB and LFIB.

3.2. Access Node Service Provisioning

AN is provisioned with a new pseudowire service instance. AN requests a required /32 FEC label from AGN1x using LDP DoD procedures.

Following the initial setup phase described in section 3.1, the first service instance gets provisioned on AN. Let us assume this is a pseudowire (PW) that is associated with either an attachment circuit (AC for VPWS service) or a virtual switching instance (VSI for VPLS service). The type of PW service does not matter. This PW will be signaled using targeted LDP FEC128 (0x80). Hence the PW is provisioned with the PW ID and the loopback IP address of destination node.

From IP/MPLS perspective, following label operations need to complete successfully to establish PW service:

- o LSP label for destination /32 FEC needs to be signaled using LDP DoD.
- o PW label for specific PW ID FEC128 needs to be signaled using targeted LDP and PWE3 signaling procedure as per RFC 4447 [RFC4447]

AN has to establish a TCP/IP connection to the destination node for the targeted LDP session. This is done either by an explicit targeted LDP session configuration on AN (most likely) or automatically at the time of provisioning the PW.

Destination node may be located in the local metro region or in a remote region. In the former case destination node is reachable via both native IP and MPLS LSPs, in the latter only via MPLS LSPs as transit core nodes do not hold remote AN IP routes. To ensure a common behavior for both cases, it is required that IP packets associated with this tLDP TCP/IP connection must be forwarded over an MPLS LSP to the destination, in other words LDP DoD label must be pushed on those packets by AN. This requires that LDP DoD is used for setting up MPLS LSP before tLDP session is established.

To establish an LSP for destination /32 FEC, AN looks up its local routing table for this /32 and chooses outgoing interface. If label for this /32 route is not already installed based on the configured static route with LDP DoD request policy, AN MUST send an LDP DoD label mapping request over this interface to adjacent AGN1. AGN1 replies with its label for this FEC. AGN1 MUST install this incoming

label in its LIB and FIB. Upon receiving label mapping Access Node MUST accept this label based on the exact route match between advertised FEC and route entry in its RIB or based on the longest match based on [RFC5283]. If AN accepts the label it MUST install it as outgoing label in its LIB and FIB.

If AN is dual homed to two AGN1's and routing entries for these AGN1's are configured as equal cost paths, Access Node MUST send LDP DoD label requests to both AGN1's and install all received labels in its LIB and FIB. If AN has multiple parallel links to AGN1 and routing entries for these links are configured as equal cost paths, same label should be used in LIB and programmed in the FIB for all these links.

Following establishment of targeted LDP session, AN and the destination node exchange their PW label bindings based on the configured PW ID, and activate the PW service.

In order to forward payload packets over the established PW, AN has to push PW encapsulation including the PW labels, followed by pushing the LSP label. AN chooses the LSP label based on the locally configured static route. If a specific route is reachable via multiple interfaces to AGN1 nodes (AN dual-homed, parallel links or both) and the route has multiple equal cost paths, Access Node MUST implement Equal Cost Multi-Path (ECMP) functionality. This involves AN to use hash-based load-balancing mechanism and send the PW packets in a flow-aware manner with appropriate LSP labels via all equal cost uplinks.

ECMP mechanism is applicable in an equal manner for parallel links between two network elements and multiple paths towards the destination. The traffic demand is distributed over the available paths.

To handle local link or adjacent node failures, AN should handle these local failures in a local-repair manner, that is AN should implement a simple LFA scheme. This will involve AN in case of primary interface failure choosing ECMP alternative or if not available a second best link.

3.3. Access Node Service Decommissioning

The last PW service instance is decommissioned. The LSP label is released.

With the decommissioning of the service, the Pseudowire is deleted. If it is the last PW to the specific destination node, targeted LDP session is not longer needed and SHOULD be terminated (automated or

by configuration).

The LSP label is not longer required on AN for carrying any service.

If the LSP label was originally requested based on the static route configuration with LDP DoD request policy, the label MUST be retained by AN.

If /32 FEC label was originally requested based on tLDP session configuration, AN SHOULD delete the label from its LIB and FIB. The deletion of the label MAY be done immediately or with a Garbage Collection mechanism.

If AN deletes the label, it MUST signal to AGN1 the Label Release Message, indicating the label is not longer required. This MAY be done immediately or with a Garbage Collection mechanism. The AGN1 MAY use this message to delete the label in its FIB, if it is not needed for other peers. However AGN1 MAY retain the label in its LIB.

3.4. Service Failure

A service instance has failed due to a network event. No impact on LDP DoD /32 FEC label.

Variety of network events can trigger PW failure:

- o Local or remote attachment circuit has failed (remote end status signaled by targeted LDP).
- o Local or remote PSN-facing PW (ingress or egress) has failed (remote end signaled by targeted LDP).
- o PW is not in a enabled state for operation.
- o PW OAM has signaled a failure (e.g. VCCV).
- o Targeted LDP session is broken.
- o Network link or node failures (described in <sec 3.5>).

In all cases except the last one, the status of the PW service MUST not have any impact on the LSP label signaling. Therefore AN MUST NOT modify associated LSP label entries in its LIB and FIB.

3.5. Network Transport Failure

Number of different network events can impact services on AN. Following sections describe network event types that impact LDP DoD operation on AN and AGN1.

3.5.1. Access Node Failure

Event: Access Node fails. Adjacent AGN1s delete LDP DoD /32 FEC labels for this AN.

If AN fails, the link between AN and AGN1 goes down.

AGN1 MUST remove associated static route(s) pointing to this AN from its routing table. AGN1 MUST also remove the associated outgoing label(s) for this AN's /32 loopback(s) from its FIB. AGN1 MAY remove the incoming labels provided to affected AN subject to other nodes using those labels or label retention procedures implemented on this AGN1.

AGN1 SHOULD implement all relevant global-repair IP/MPLS procedures to propagate the AN failure towards the network.

3.5.2. Access Node Uplink Failure

Event: Link between AN and AGN1 fails. If last link, LDP DoD /32 FEC labels get deleted on AN and AGN1s.

In the event when AN-AGN1 link fails (and there are no more active L3 links connecting AN and AGN1) or the LDP DoD session between AN and AGN1 fails, both AN and AGN1 should take following actions.

- o AN MUST delete /32 FEC label entries for labels provided by AGN1 affected by this link failure event, and remove those labels from its LIB and FIB tables.
- o AGN1 MUST remove associated static route(s) pointing to this AN from its routing table.
- o AGN1 MUST also remove the associated outgoing label(s) for this AN's /32 loopback(s) from its FIB.
- o AGN1 MAY remove the incoming labels provided to affected AN subject to other nodes using those labels or label retention procedures implemented on this AGN1.
- o In the event of interface or LDP DoD session coming back up, AN MUST request required labels observing procedures against link

flapping.

- o If AN is not redundantly connected to AGN1 nodes, in case of link failure, the service MUST go into a failure status on AN. When the failed link comes up again, the service MUST be reestablished automatically.
- o If AN is connected to AGN1 over multiple parallel links implementing ECMP, in case of link failure, the service SHOULD be immediately re-routed to remaining links with the same /32 FEC label.
- o If AN is dual homed to two AGN1s (i.e. AGN11 and AGN12 in the reference topology), in case of link failure and isolation from one of the AGN1s, the service SHOULD be immediately re-routed by AN to use link(s) to the other AGN1.

3.5.3. AGN node failure

AGN1 node fails. Adjacent ANs delete LDP DoD /32 FEC labels provided by this AGN1.

If AGN1 fails, all links between this AGN1 and adjacent ANs go down.

AN MUST remove from its RIB static route(s) pointing to this AGN1 as a next hop. AN MUST also remove from its LIB and FIB all the outgoing labels provided by now failed AGN1. Access Node SHOULD use local-repair procedures to re-route away from the failure.

3.5.4. AGN network-side reachability failure

AGN1 loses network reachability to a specific destination or set of destinations. Associated /32 FEC labels are withdrawn from AN.

In case of a network event that makes AGN1 lose its network reachability to a destination or set of destinations used by AN, AGN1 MUST send LDP Label Withdraw messages to all local ANs and withdraw labels for affected /32 FECs. Upon receiving those messages from AGN1, ANs MUST remove those labels from their LIB and FIB tables, and use alternative LSPs instead if available as part of global-repair.

4. LDP Downstream on Demand Procedures

Label Distribution Protocol is specified in RFC5036 [RFC5036], and all LDP DoD implementations must follow this specification.

In MPLS network traffic flows from upstream to downstream LSR

(RFC3031 [RFC3031], section 3.2). In case of downstream assigned labels as described in this document, labels are assigned by the downstream LSR and signaled to the upstream LSR as shown in figure below.

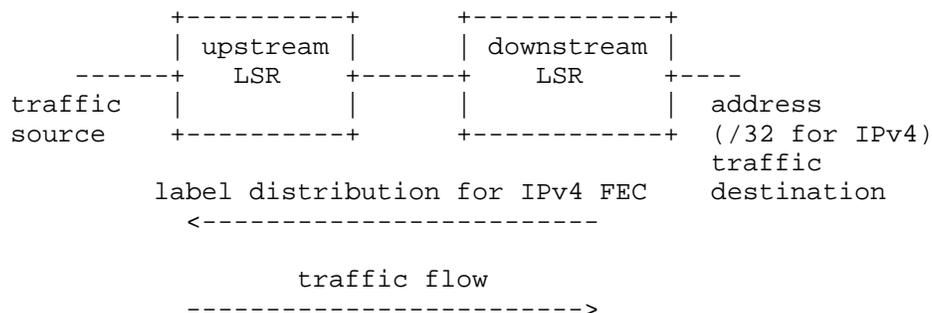


Figure 4. LDP label assignment direction

4.1. LDP Label Distribution Modes

The LDP protocol specification RFC5036 [RFC5036] section 2.6) defines two modes for label distribution control:

- o Independent - an LSR recognizes a particular FEC and makes the decision to bind a label to the FEC independently to distribute the bindings to its peers. The new FECs are recognized whenever new routes become visible to the router.
- o Ordered - an LSR binds a label to a particular FEC if and only if it is the egress router or it has received a label binding for the FEC from its next hop LSR.

The LDP protocol specification (RFC5036 [RFC5036] section 2.6) defines two modes for label retention:

- o Conservative - the bindings between a label and an FEC received from LSRs that are not the next hop for a given FEC are discarded. This mode requires an LSR to maintain fewer labels.
- o Liberal - the bindings between a label and an FEC received from LSRs that are not the next hop for a given FEC are retained. This mode allows for quicker adaptation to topology changes and allows for the switching of traffic to other LSRs in case of change.

Note: In conservative label retention mode, if the next hop for FEC changes, then the LSR has to request a new label from the new next hop before labeled packets can be forwarded.

For the LDP DoD Advertisement mode on AN an ordered label distribution mode and conservative label retention mode MUST be supported.

With the ordered distribution mode, the AGN1 provides the access node only with FEC/label binding, where the AGN1 has a correct label binding itself.

With the conservative retention mode, the AN retains only FEC/label binding on an interface, if the interface where the FEC/label mapping was received is a valid next hop. The DoD mode is explicitly mentioned in the description of the conservative mode in (RFC5036 [RFC5036] section 2.6.2.1).

The downstream labels on AGN1 may be allocated by LDP Downstream on Demand (LDP DoD) or LDP Downstream Unsolicited (LDP DU) or BGP labeled unicast routes that are learned as per RFC 3107 [RFC3107]. AGN1 should use the conservative label retention mode in case of Downstream on Demand label advertisement. AGN1 should use liberal retention mode in case of LDP DU label advertisement mode. AGN1 should establish either LDP DoD or LDP DU session to peer LSR based on LDP session negotiation procedure, specified in section 4.3. AGN1 must support interworking between LDP DoD and LDP DU sessions to different LSR peers.

4.2. IPv6 Support

The current standard specifies the usage of IPv4 in LDP either as transport protocol or as service (binding of MPLS labels to Ipv4 addresses). RFC5036 [RFC5036] also describes the usage of IPv6 as transport protocol, but not as the service. For the future deployment, LDP DoD MUST also support IPv6 for transport and services. This is still under development ([I-D.ietf-mpls-ldp-ipv6]).

4.3. LDP DoD Session Negotiation

The AN should propose the Downstream on Demand label advertisement by setting "A" value as 1 in the Common Session Parameters TLV of the Initialization message. The rules for negotiating the label advertisement mode are specified in the section 3.5.3 of LDP protocol specification (RFC5036 [RFC5036]).

To establish Downstream on Demand session both AN and AGN1 should propose the Downstream on Demand label advertisement mode in the Initialization message for other than ATM/FR links. If AGN1 proposes Downstream Unsolicited mode, AN should send Notification with status "Session Rejected/Parameters Advertisement Mode" and then close the

LDP session.

If AN is acting as active role, it should re-attempt the LDP session immediately. If AN receives same Downstream Unsolicited mode again, AN should follow the exponential backoff algorithm as specified in the (RFC5036 [RFC5036] with delay of 15 seconds and subsequent delays grow to a maximum delay of 2 minutes.

In case a PWE3 service is required between AN and AGN1, and LDP DoD has been negotiated for IPv4 and IPv6 FECs, the same LDP session should be used for PWE3 FECs. Even if DoD label advertisement has been negotiated for IPv4 and IPv6 LDP FECs as described earlier, LDP session should use Downstream Unsolicited label advertisement for PWE3 FECs as specified in RFC4447 [RFC4447].

4.4. Label Request Procedures

The access node requests an MPLS label from the AGN1 with the Label Request Message (RFC5036 [RFC5036] , section 3.5.8). The FEC is the specific IP address of the requested Forwarding Equivalent Class (FEC). The MPLS Label is delivered with the Label Mapping Message (RFC5036 [RFC5036] section 3.5.7).

4.4.1. AN Label Request Procedure

AN will request label bindings for AGN1 nodes as well as labels for the possible loopback addresses within seamless MPLS network based on following trigger events in addition to RFC5036 [RFC5036], section 3.5.8.1:

- o AN is configured with /32 static route with LDP DoD label request policy in line with AN start-up use case specified in section 3.1.
- o AN is configured with service (e.g. PW) in line with PW provisioning use case described in section 2.2.
- o AN and AGN1 link comes up and LDP DoD session is established. In this case AN should send label request messages for all /32 static routes configured with LDP DoD policy and all /32 routes related to provisioned services (PW) that are not covered by static routes with LDP DoD policy.

AGN1 will respond with label mapping message with a non-null label if any of the below conditions are met on AGN1:

- o Requested FEC is a BGP labelled unicast route RFC 3107 [RFC3107] and this BGP route is the best selected for this FEC or

- o Requested FEC is an IGP or static route and there is an LDP label already learnt from downstream router (by LDP DU or LDP DoD), and this downstream router is the best next hop selected for this FEC. If selected downstream peer is LDP DoD and there is no label for this FEC, AGN1 will send a further label request message to this peer. In such case AGN1 will respond to AN only after getting a label from downstream peer.

AGN1 may send a label mapping with explicit-null or implicit-null label if it is acting as an egress for the requested FEC, or it may respond with "No Route" notification if no route exists.

4.4.2. AGN Label Request Procedure

AGN should send label request message based on the following trigger events in addition to RFC5036 [RFC5036], section 3.5.8.1:

- o AGN receives a label request from upstream LDP peer, has no downstream label for requested FEC and the downstream peer is LDP DoD, or
- o AGN is configured with /32 static route with LDP DoD label request policy.

In case of ECMP, AGN should send label requests over all LDP DoD sessions associated with selected ECMP best next hops. In case of LFA, AGN should request labels over LDP DoD sessions associated with both primary and backup next hop routers.

In both ECMP and LFA cases, downstream LSR may be AN. AN should respond back with label mapping to AGN if corresponding /32 route configuration (loopback address) exists, otherwise AN responds with "No route" notification.

4.4.3. Label Request Retry Procedure

If AN or AGN receives a "No route" Notification in response to its label request message, it should retry with exponential backoff algorithm similar to the backoff algorithm mentioned in the LDP session negotiation section 4.3.

If there is no response to the sent label request message, the LDP specification RFC 5036 [RFC5036] (section A.1.1, page# 100) states that LSR should not send another request for the same label to the peer and mandates that a duplicate label request is considered a protocol error and should be dropped by the receiving LSR by sending Notification message.

AN or AGN1 should not send duplicate label request message again if there is no response from downstream peer.

If the static route gets deleted or DoD request policy rejected for particular FEC before receiving label mapping message, then AN or AGN1 should send a Label Abort message to downstream router.

4.5. Label Withdraw Procedure

If an MPLS label in the AGN1 is no longer valid, the AGN1 withdraws this FEC/label binding from the access nodes with the Label Withdraw Message (RFC 5036 [RFC5036] section 3.5.10) with a specified label TLV or with an empty label TLV.

AGN1 should withdraw a label for specific FEC in the following cases:

- o If LDP DoD ingress label is associated with an outgoing label assigned by BGP labelled unicast route, this route is withdrawn.
- o If LDP DoD ingress label is associated with an outgoing label assigned by LDP (DoD or DU) and IGP route is withdrawn from the RIB or downstream LDP session is lost.
- o If LDP DoD ingress label is associated with an outgoing label assigned by LDP (DoD or DU) and received label is withdrawn by the downstream LSR.
- o If LDP DoD ingress label is associated with an outgoing label assigned by LDP, route next hop changed and
 - A. there is no LDP session to the new next hop. To minimize probability of this, AGN should implement LDP-IGP synchronization procedures as specified in RFC 5443 [RFC5443].
 - B. there is an LDP session but no label from downstream LSR. See note below.
- o If AGN1 is configured with a policy to reject exporting label mappings to AN.

The access node responds with the Label Release Message (RFC5036 [RFC5036] section 3.5.11).

After sending label release message to AGN1, AN should keep retry sending label request message again, assuming AN still requires the label.

AN should withdraw a label if the local route configuration (/32

loopback) is deleted on the AN. But if a service (PW) is decommissioned, then AN will release the label - see cases described in the next section 4.6.

Note: For any events inducing next hop change, AGN1 should attempt to converge the LSP locally before withdrawing the label from an upstream AN. For example if the next hop changes for a particular FEC and if the new next hop allocates labels by LDP DoD session, then AGN1 must send label request on the new next hop session. If AGN1 doesn't get label mapping for some duration, then and only then AGN1 must withdraw the upstream label.

4.6. Label Release Procedure

If an access node does not need any longer a label for an FEC, it sends a Label Release Message (RFC 5036 [RFC5036] section 3.5.11) to the AGN1 with or without the label TLV.

If AN or AGN1 receive unsolicited label mapping on DoD session, they should release the label by sending label release message.

AN should send a label release message in the following cases:

- o If it receives a label withdraw from AGN.
- o If /32 static route with LDP DoD label request policy is deleted.
- o If service (PW) gets decommissioned and there is no corresponding /32 static route with LDP DoD label request policy configured.
- o If the route next hop changed, and the label does not point to the best next hop.

AGN should send a label release message to downstream DoD session in the following cases:

- o If /32 static route with LDP DoD label request policy is deleted.
- o If the route next hop changed, then AGN should send label release on the old next hop DoD session.
- o If it receives label withdraw from downstream DoD session.

4.7. Local Repair

To support local-repair with LFA, AGN1 should request labels from both primary (best) next hop and for backup (second best) next hop LDP DoD sessions as specified in the label request procedures in the

section 4.4.2. This will enable AN1 to pre-program the backup forwarding path with the backup label if the primary next hop link fails, and invoke LFA switch-over procedure.

To support local repair on AN, AN should request label from the backup (second best) next hop. This will enable AN1 to pre-program the backup forwarding path with the backup label if the primary next hop link fails, and invoke LFA switch-over procedure.

5. IANA Considerations

This document makes no request of IANA.

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

6. Security Considerations

7. Acknowledgements

The authors would like to thank Nischal Sheth, Nitin Bahadur and Nicolai Leymann for their suggestions and review.

8. References

8.1. Normative References

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

8.2. Informative References

[I-D.ietf-mpls-ldp-ipv6]
Manral, V., Papneja, R., Asati, R., and C. Pignataro,
"Updates to LDP for IPv6", draft-ietf-mpls-ldp-ipv6-04
(work in progress), May 2011.

[I-D.ietf-mpls-seamless-mpls]
Leymann, N., Decraene, B., Filsfils, C., Konstantynowicz,
M., and D. Steinberg, "Seamless MPLS Architecture",
draft-ietf-mpls-seamless-mpls-00 (work in progress),
May 2011.

[RFC3031] Rosen, E., Viswanathan, A., and R. Callon, "Multiprotocol

Label Switching Architecture", RFC 3031, January 2001.

- [RFC3107] Rekhter, Y. and E. Rosen, "Carrying Label Information in BGP-4", RFC 3107, May 2001.
- [RFC4446] Martini, L., "IANA Allocations for Pseudowire Edge to Edge Emulation (PWE3)", BCP 116, RFC 4446, April 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.
- [RFC5036] Andersson, L., Minei, I., and B. Thomas, "LDP Specification", RFC 5036, October 2007.
- [RFC5283] Decraene, B., Le Roux, J.L., and I. Minei, "LDP Extension for Inter-Area Label Switched Paths (LSPs)", RFC 5283, July 2008.
- [RFC5443] Jork, M., Atlas, A., and L. Fang, "LDP IGP Synchronization", RFC 5443, March 2009.

Authors' Addresses

Thomas Beckhaus
Deutsche Telekom AG
Heinrich-Hertz-Strasse 3-7
Darmstadt, 64307
Germany

Phone: +49 6151 58 12825
Fax:
Email: thomas.beckhaus@telekom.de
URI:

Bruno Decraene
France Telecom
38-40 rue du General Leclerc
Issy Moulineaux cedex 9, 92794
France

Phone:
Fax:
Email: bruno.decraene@orange-ftgroup.com
URI:

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

Phone: 1-(978)-589-8861
Fax:
Email: kishoret@juniper.net
URI:

Maciek Konstantynowicz
Juniper Networks

Phone:
Fax:
Email: maciek@juniper.net
URI:

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: November 20, 2011

M. Bhatia
Alcatel-Lucent
L. Jin
ZTE
F. Jounay
France Telecom
May 19, 2011

Extensions to Resource Reservation Protocol - Traffic Engineering
(RSVP-TE) for Bi-directional Label Switched Paths (LSPs)
draft-bhatia-mpls-rsvp-te-bidirectional-lsp-01

Abstract

There are several applications that require symmetric Multiprotocol Label Switching (MPLS) path between two points. This cannot be achieved with regular MPLS as the LSPs are unidirectional. If symmetry is required, a separate LSP in each direction is required for bidirectional traffic flow. Generalized MPLS on the other hand, has provisions for setting up a bidirectional LSP. This document uses the extensions introduced for GMPLS and applies it to regular MPLS for establishing bidirectional LSPs. Additionally, it also describes how bi-directional symmetrical Fast Reroute using both one-to-one and facility backup can be achieved.

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].

When used in lower case, these words convey their typical use in common language, and are not to be interpreted as described in RFC2119 [RFC2119].

Status of this Memo

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

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

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

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

This Internet-Draft will expire on November 20, 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	4
2. RSVP-TE to signal Bi-directional LSP	4
3. Fast Reroute mechanisms	5
3.1. Discovering Upstream Labels	6
3.2. Failure detection between PLR and MP	7
4. Behavior of various network elements in FRR	7
4.1. The Head-End Router Behavior	7
4.2. The Point of Local Repair (PLR) Behavior	8
4.2.1. PLR Behavior during one-to-one backup for a node failure	9
4.2.2. PLR Behavior during facility backup for a node failure	10
4.3. The Merge Point (MP) Router Behavior	11
5. Security Considerations	13
6. IANA Considerations	13
7. References	13
7.1. Normative References	13
7.2. Informative References	13
Authors' Addresses	14

1. Introduction

There are several applications that require symmetrical paths between a pair of speakers. One such application is 1588 [IEEE-1588] which requires that the Delay_Resp message takes the same path as the associated Delay_Req message. [I-D.ietf-tictoc-1588overmpls] describes a method for transporting PTP messages (PDUs) over an MPLS network to enable proper handling of these packets. Currently, the only way to ensure that the different PTP messages follow a symmetrical path is by statically configuring the RSVP-TE LSPs. This is unscalable and will not work in case of network failures as MPLS FRR may not guarantee symmetrical alternate paths.

This document describes how RSVP-TE can be used for setting up bi-directional LSPs for regular MPLS and the extensions required in FRR to ensure that the alternate paths are also symmetrical.

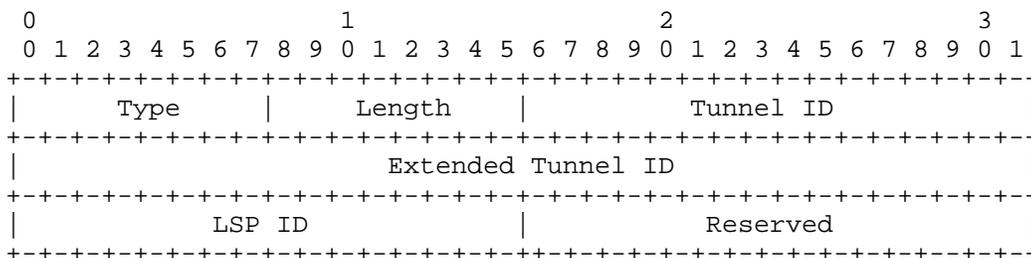
2. RSVP-TE to signal Bi-directional LSP

[RFC3473] describes a point-to-point bidirectional LSP mechanism for the GMPLS architecture, where a bidirectional LSP setup is indicated by the presence of an Upstream Label in the Path message. The Upstream_Label object has the same format as the generalized label, and uses Class-Number 35 (of form 0bbbbbbb) and the C-Type of the label being used.

For regular MPLS the Upstream_Label object will be used with C-Type value of 1.

Typically, a node initiates an RSVP session by adding the RRO to the Path message. The initial RRO contains only one subobject - the sender's IP addresses. If the node also desires label recording, it sets the Label_Recording flag in the SESSION_ATTRIBUTE object. This document extends this mechanism by also adding the Upstream label that has been advertised in the RRO subobject. Thus the initial RRO will now contain the sender's IP address and the Upstream label advertised by it. The upstream label subobject in RRO object will be with type 0x04 and same C-type with label object.

It is necessary to ensure the PLR and MP to bind to the same bidirectional protection tunnel (bypass tunnel or detour tunnel), this draft introduces a new subobject in RRO object to indicate the tunnel that PLR or MP binds.



Type 0x05 Protection bidirectional tunnel ID

Length The Length contains the total length of the subobject in bytes, including the Type and Length fields. The Length is always 8.

The tunnel ID and extended tunnel ID is derived from session object, LSP ID is derived from sender_template object of protection LSP.

3. Fast Reroute mechanisms

[RFC4090] extensions can be used to perform fast reroute for the mechanism described in this document when applied within packet networks. This section only applies to LSRs that support [RFC4090].

This section uses terminology defined in [RFC4090], and fast reroute procedures defined in [RFC4090] MUST be followed unless specified below. The head-end and transit LSRs MUST follow the SESSION_ATTRIBUTE and FAST_REROUTE object processing as specified in [RFC4090] for each Path message.

Since its a bi-directional LSP the detour LSPs and the bypass tunnels that are used for the protected LSP must also be bi-directional. This is required so that path symmetry is maintained even in an event of a network failure.

It should be noted that in case of bi-directional LSPs, the LSRs involved will play the role of both the Point-of-Local-Repair (PLR) and Merge Point (MP) at the same time during the failure. The router that is the PLR will become the MP for the traffic thats coming from the opposite direction.

In the Figure 1 assume that ABCD is the protected LSP. For protecting link BC, there is a bidirectional bypass tunnel BEFC (or a detour LSP in case of 1-on-1). B is the PLR and C is the MP for the traffic flowing from A towards D and B is the MP, and C the PLR for traffic flowing in the opposite direction (from D towards A).

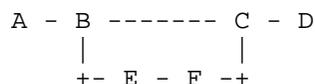


Fig 1: Topology for
link protection

In the Figure 2 ABCDE is the protected LSP and BFGD is the bypass tunnel for protecting the node C. In this case B is the PLR and D the MP for traffic from A towards E, and the roles reverse, i.e. B becomes the MP and D the PLR for traffic flowing in the opposite direction (from E towards A).

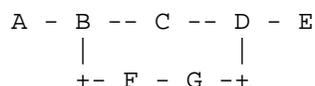


Fig 2: Topology for
node protection

3.1. Discovering Upstream Labels

To support facility backup, the PLR must determine a label that will indicate to the MP that packets received with that label should be switched along the protected LSP. This can be done without explicitly signaling the backup path if the MP uses a label space global to that LSR.

As described in [RFC4090], the head-end LSR MUST set the "label recording requested" flag in the SESSION_ATTRIBUTE object for LSPs requesting local protection. This will cause (as specified in [RFC3209]) all LSRs to record their INBOUND labels and to note via a flag whether the label is global to the LSR. Thus, when a protected LSP is first signaled through a PLR, the PLR can examine the RRO in the Resv message and learn about the incoming labels that are used by all downstream nodes for this LSP. Similarly the MP, which will become the PLR for the reverse direction will learn about the upstream labels that are being used by the upstream nodes for this LSP by examining the upstream label subobject in RRO in the Path message.

The bypass tunnels and the detour tunnels that are set up for a bidirectional LSP must be bidirectional as well. They can internally use the Upstream_label technique that was described earlier to establish a bidirectional LSP.

3.2. Failure detection between PLR and MP

It is required that PLR and MP should detect the failure at the same time, then the two nodes could switch with traffic to the protection tunnel (bypass tunnel or detour tunnel) simultaneously. Such kind of detection mechanism could be BFD [RFC5880], RSVP-TE hello, or other proper mechanism.

For the link protection scenario, the detection mechanism should be enabled between PLR and MP. When a failure happens, both PLR and MP could detect the failure simultaneously, and switch the traffic to the protection tunnel.

For the node protection scenario, it is required to setup two correlated detection sessions. For the figure2 topology in section 3, the PLR node B and MP node D will do the node protection for the protected tunnel. There will be a detection session1 on the link between B and C, and session2 between C and D.. When a link failure happens between B and C, B could detect the failure by the session1, C should notify the link failure event to D by setting the diagnostic code to 6 (Concatenated Path Down) in BFD control packet [RFC5880]. Then D could detect failure through BFD control packets in session 2.

An alternative way is to do protected LSP segment detection between PLR and MP. When the link or node failed, the protected LSP segment detection session will be down, and both PLR and MP could detect the failure.

4. Behavior of various network elements in FRR

When a failure happens in the network the PLR router closest to the failure must perform the traffic protection. The MP router is the router that is the next hop to the failure point and merges the protected traffic back to the original path. In case of bidirectional LSPs, the same LSR is PLR in one direction and the MP for the other. Let us examine in detail what each network element does for the MPLS FRR.

4.1. The Head-End Router Behavior

The Head-End router originates the bi-directional LSP that needs to be protected. It's here that the desired protection type (one-to-one or facility backup) is also defined.

The Path message which has the FRR information in the SESSION_ATTRIBUTE object is propagated from the head-end LSR to the Tail router. Each hop sees the FRR flags and assumes the PLR role

and tries to establish a bi-directional tunnel. Every hop reports the availability of the FRR protection if its able to establish a bi-directional tunnel successfully. This is done via setting the RRO flags in the Resv message.

When a network failure occurs the PLR, or router upstream of the failure to be precise, uses FRR to reroute the traffic around the failure, and notifies the head-end LSR by (i) setting the FRR "Local protection in use" flag (0x2) in the RRO object of the Resv message and (ii) by sending a PathErr message with an ERROR object with code 0x19 - RSVP Notify Error and error value 0x3 - Tunnel locally repaired. The router that is downstream of the failure (traditionally the MP in case of unidirectional LSPs) also uses FRR to reroute the traffic around the failure. It does not send any message to the head-end LSR.

The head-end LSR upon receiving this indication tries to switch the traffic to a secondary LSP if its available. In case its not active, the head-end LSR signals this LSP via make-before-break mechanism.

4.2. The Point of Local Repair (PLR) Behavior

The PLR router of the protected LSP is also the origination point (head-end Router) of the protection tunnel (detour LSP or bypass tunnel). It is also the MP for the reverse protection tunnel at the same time. When an intermediate LSR receives a Path message carrying a SESSION_ATTRIBUTE with the FRR flags set, it assumes the role of a PLR and starts signaling a bi-directional FRR protection tunnel. In case facility backup is requested by the head-end LSR, the PLR signals a new bi-directional tunnel only if a bypass tunnel fulfilling the requirements does not already exist.

In the sections that follow the terms upstream and downstream are used in reference to the direction of traffic flow from the head-end towards the tail end. Thus router the tail router is downstream to the head-end.

When a network failure happens, the upstream router local to the failure assumes the role of the PLR and switches the traffic to the protection tunnel. This PLR is from now on referred to as the "upstream PLR". The downstream router, local to the failure also assumes the role of the PLR and switches the traffic to the bidirectional protection tunnel that is set up. This PLR is referred to as the "downstream PLR". These routers can use either the bi-directional detour LSP or a bi-directional bypass tunnel, depending upon what was requested by the head-end LSR.

The egress label that each PLR uses depends upon the kind of

protection provided. The subsequent sections only describe the behavior of the "upstream PLR" that is different with the protection mechanisms as described in [RFC4090].

Once the traffic gets switched to the protection path, the "downstream" PLR does not need to inform the HE router about the network failure.

4.2.1. PLR Behavior during one-to-one backup for a node failure

For the one-to-one backup, MP should bind the backup tunnel to protected LSP before replying the RESV message of detour LSP. When the PLR setup the detour LSP and bind to the protected LSP successfully, that also indicates that MP has bound successfully.

In case of one-to-one backup, the protection or the detour tunnel is a regular LSP. The downstream PLR uses the label that was distributed by the immediate upstream router on the detour LSP (detour label) to detour traffic arriving from the downstream router of the protected LSP. The label arriving from the immediate downstream router of the protected tunnel is swapped with the detour label, and the traffic is sent through the detour LSP.

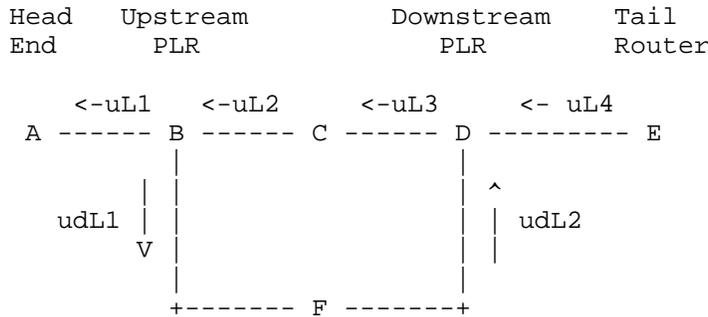


Fig 3: one-to-one FRR protection

ABCDE is a bi-directional protected LSP
 BFD is a bi-directional detour LSP

The above figure describes this mechanism. udL1 is the Upstream_label advertised by B when setting up the bi-directional detour LSP from B to D. Similarly, udL2 is the Upstream_label advertised by F to D, when setting up this LSP. uL1, uL2, uL3 and uL4 are the Upstream_labels advertised when setting up the bi-directional protected tunnel ABCDE.

When a network failure happens, in this case the LSR router between B and D fails, the node D will assume the role of a downstream PLR and would need to switch the traffic from the protected LSP to the detour LSP. D does this by programming a Swap operation on the egress of the protected LSP path to the egress of the detour LSP. The uL4 label is thus swapped with udL2 during the failure, instead of label uL3.

4.2.2. PLR Behavior during facility backup for a node failure

For the facility backup, when the PLR successfully bind the protection tunnel to the protected LSP, it SHOULD insert the Protection Tunnel subobject in RRO object in the path message, and send downstream.

In the case of facility backup, the data from the protected LSP is tunneled through the bypass tunnel. Therefore, the outer label of the tunneled packet in the reverse direction is the label distributed by the immediate upstream router of the bypass tunnel. The "downstream PLR" also needs to know what label was expected by the router where this tunneled traffic merges (MP) at the upstream. The record label option makes this information available from the RRO in the Path messages for the protected LSP. This is the inner label that must be in the tunneled packet. Thus, the "downstream PLR" swaps the incoming label from the immediate downstream router in the protected path with these two labels and sends the path through the bypass tunnel.

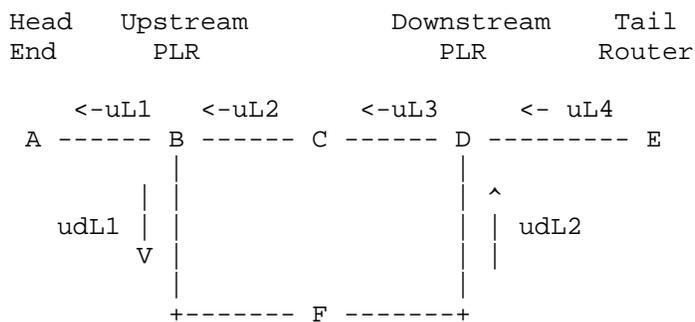


Fig 4: Facility backup protection

ABCDE is a bi-directional protected LSP
 BFD is a bi-directional bypass tunnel

The above figure describes the topology and the labels exchanged.

udL1 is the Upstream_label advertised by B when setting up the bi-directional facility bypass tunnel from B to D. Similarly, udL2 is the Upstream_label advertised by F to D, when setting up this tunnel. uL1, uL2, uL3 and uL4 are the Upstream_labels advertised when setting up the bi-directional protected tunnel ABCDE.

The "downstream" PLR router (D in this case) knows the label (uL2 in this case) that the upstream NNHop router expects because it has received a Path message which had this Upstream_label recorded in the RRO.

When a network failure happens, in this case the LSR router between B and D fails, the node D will assume the role of a "downstream" PLR and reroutes the traffic from the protected LSP through the bypass tunnel as follows:

- o PLR D performs a swap operation to change the transport label. Since it knows that its doing node protection over the bypass tunnel, it will use the label that the NNHop router ("upstream" MP) expects instead of the label that the Nhop router (failed LSR) expects. D thus, swaps out uL4 and replaces it with uL2, instead of uL3 as it would normally have done.
- o D also pushes the label udL2 on top of the label stack. This label would be used to switch the packet on the bypass tunnel and would finally reach the MP, which happens to be B in our case.

4.3. The Merge Point (MP) Router Behavior

The MP router is the LSR where the protection tunnel (detour LSP or bypass tunnel) and the protected LSP meet. It is the termination point (Tail router) of the protection tunnel. For a bi-directional protection tunnel the MP router in one direction becomes the PLR in the other.

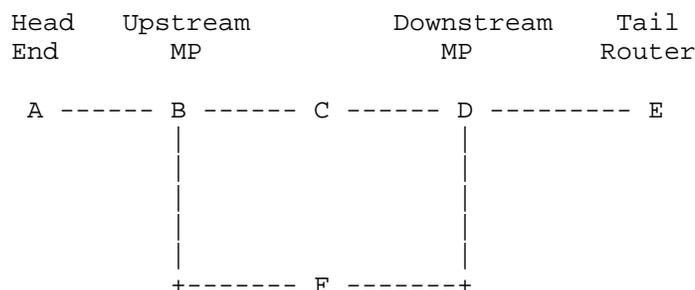


Fig 5: Merge Points in bi-directional FRR

ABCDE is a bi-directional protected LSP
 BFD is a bi-directional protection tunnel

Figure 5 shows two MPs associated with a bi-directional protection tunnel. This document refers to the MP defined in [RFC4090] as a downstream MP. This document does not change the behavior of the downstream MP. This means that it is still responsible for maintaining the protection tunnel's state by sending the Resv messages to the PLR and is also responsible for maintaining the state of the protected tunnel during the network failure. The upstream MP, defined in this document, is not required to do any of these. Its only responsible for merging the reverse traffic back to the protected path.

In one-to-one backup, the tail-end of backup LSP should consider it as MP. When the tail-end receives the Path message, and before sending RESV, it should try to bind the backup tunnel to protected tunnel. When binding successfully, MP sends the RESV message upstream for the backup tunnel.

During one-to-one backup the MP performs a swap operation on the ingress label of bi-directional detour LSP with the egress label of the bi-directional protected LSP.

In facility protection, when the LSR receives the Path message with RRO object, indicating the Previous_Hop or Previous_Previous_Hop with Protection Tunnel subobject, it should consider itself as MP. And it SHOULD try to bind the same protection tunnel indicated by Protection Tunnel subobject to the protected LSP. The protection tunnel would be expected to be from MP to PLR, with same tunnel-ID and LSP-ID indicated by the subobject.

During facility protection, the traffic arrives with a bypass tunnel label. The MP pops out this label to expose the original protected tunnel label that was distributed to the immediate downstream router

via the Upstream_label mechanism in the Path message on the protected tunnel. Since this label is already programmed, the traffic is switched out correctly.

The Resv message from MP to PLR should be sent in the protection LSP since there is a LSP path from MP to PLR.

5. Security Considerations

This document raises no new security concerns.

6. IANA Considerations

No requests for IANA at this point of time.

7. References

7.1. Normative References

- [IEEE-1588] "IEEE Standard for a Precision Clock Synchronization Protocol for Networked Measurement and Control Systems", 2008.
- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC3473] Berger, L., "Generalized Multi-Protocol Label Switching (GMPLS) Signaling Resource ReserVation Protocol-Traffic Engineering (RSVP-TE) Extensions", RFC 3473, January 2003.
- [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.
- [RFC5467] Berger, L., Takacs, A., Caviglia, D., Fedyk, D., and J. Meuric, "GMPLS Asymmetric Bandwidth Bidirectional Label Switched Paths (LSPs)", RFC 5467, March 2009.

7.2. Informative References

- [I-D.ietf-tictoc-1588overmpls] Davari, S., Oren, A., Martini, L., Bhatia, M., and P. Roberts, "Transporting PTP messages (1588) over MPLS

Networks", draft-ietf-tictoc-1588overmpls-00 (work in progress), January 2011.

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

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

[RFC5880] Katz, D. and D. Ward, "Bidirectional Forwarding Detection (BFD)", RFC 5880, June 2010.

Authors' Addresses

Manav Bhatia
Alcatel-Lucent
India

Email: manav.bhatia@alcatel-lucent.com

Lizhong Jin
ZTE
889, Bibo Road
Shanghai, 201203, China

Email: lizhong.jin@zte.com.cn

Frederic Jounay
France Telecom
2, avenue Pierre-Marzin
22307 Lannion Cedex, FRANCE

Email: frederic.jounay@orange-ftgroup.com

Internet Engineering Task Force
Internet-Draft
Intended status: Informational
Expires: January 7, 2012

G. Chen
L. Li
China Mobile
July 6, 2011

IPv6 Provider Edge Routers (6PE) Information Base (MIB)
draft-chen-mpls-6pe-mib-02

Abstract

This memo defines a portion of the Management Information Base (MIB) for use with network management protocols in the Internet community. In particular, it describes a MIB module for IPv6 Provider Edge Routers (6PE) over Multiprotocol Label Switching (MPLS) Label Switching Routers (LSRs).

Status of this Memo

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

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

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

This Internet-Draft will expire on January 7, 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.

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

Table of Contents

1. Introduction	3
2. The Internet-Standard Management Framework	3
3. Overview of MIB objects	3
3.1. 6PETunnelIfTable	4
3.2. 6PEMplsIfTable	4
4. 6PE-MPLS-STD-MIB Module Definitions	4
5. Security Considerations	7
6. IANA Considerations	7
7. Normative References	7
Authors' Addresses	8

1. Introduction

IPv6 Provider Edge Routers (6PE) is a IPv6 transition technology, which could shift network to provide IPv6 access depending on existing Multiprotocol Label Switching (MPLS) core network. Operators could deploy IPv6 network without modifying IPv4 enable MPLS cloud. Therefore, 6PE is treated as a IPv6 transition solution on the early stage. 6PE will be adopted in more and more operational IP networks on account of IPv4 depletion and incremental advantages.

RFC 4789[RFC4789] has elaborated 6PE technology. When tunneling IPv6 packets over the IPv4 MPLS backbone, rather than successively prepend an IPv4 header and then perform label imposition based on the IPv4 header, the ingress 6PE Router MUST directly perform label imposition of the IPv6 header without prepending any IPv4 header. In respect of managing IPv6 tunnel, RFC 4087[RFC4087] has specified managed objects used for managing tunnels of any type over IPv4 and IPv6 networks. However, This MIB module does not support tunnels over non-IP networks. RFC 4382[RFC4382] has defined managed objects to configure and monitor MPLS layer 3 Virtual Private Networks. Nevertheless, 6PE is neither Layer 3 IP tunnel nor MPLS layer 3 VPN. This document is aimed at discribing managed objects for 6PE.

2. The Internet-Standard Management Framework

For a detailed overview of the documents that describe the current Internet-Standard Management Framework, please refer to section 7 of RFC 3410[RFC3410] .

Managed objects are accessed via a virtual information store, termed the Management Information Base or MIB. MIB objects are generally accessed through the Simple Network Management Protocol (SNMP). Objects in the MIB are defined using the mechanisms defined in the Structure of Management Information (SMI). This memo specifies a MIB module that is compliant to the SMIV2, which is described in STD 58, RFC 2578[RFC2578], STD 58, RFC 2579[RFC2579] and STD 58, RFC 2580[RFC2580]

3. Overview of MIB objects

The following subsections describe the purpose of each of the objects contained in the 6PE-MPLS-STD-MIB.

3.1. 6PETunnelIfTable

6PETunnelIfTable are defined in the MIBs defining the encapsulation. An entry in the 6PE Tunnel MIB will exist for every interface entry with this interface type. An implementation of the 6PE Tunnel MIB may allow 6PETunnelIfTable to be created. Creating a tunnel will also add an entry in the 6PETunnelIfTable, and deleting a tunnel will likewise delete the entry in the 6PETunnelIfTable.

3.2. 6PEmplsIfTable

This table controls MPLS-specific parameters when 6PE is going to be carried over MPLS cloud.

4. 6PE-MPLS-STD-MIB Module Definitions

IMPORTS

MODULE-IDENTITY, OBJECT-TYPE, transmission,

Integer32, IpAddress FROM SNMPv2-SMI -- [RFC2578]

RowStatus, StorageType FROM SNMPv2-TC -- [RFC2579]

MODULE-COMPLIANCE,

OBJECT-GROUP FROM SNMPv2-CONF -- [RFC2580]

InetAddressType,

InetAddress FROM INET-ADDRESS-MIB -- [RFC4001]

ifIndex,

InterfaceIndexOrZero FROM IF-MIB -- [RFC2863]

MplsTunnelIndex, MplsTunnelInstanceIndex,

MplsLdpIdentifier, MplsLsrIdentifier

FROM MPLS-TC-STD-MIB -- [RFC3811]

MplsIndexType

FROM MPLS-LSR-STD-MIB -- [RFC3813]

```
6peMplsStdMIB MODULE-IDENTITY
LAST-UPDATED "201107060000Z" -- 06 July 2011 00:00:00 GMT
ORGANIZATION "IPv6 Provider Edge Routers (6PE) Working Group."
CONTACT-INFO
"
    Chen Gang, Editor
    Email: chengang@chinamobile.com

    Li Lianyuan, Editor
    Email: lilianyuan@chinamobile.com
"
DESCRIPTION
    "This MIB module complements the 6PE-MPLS-STD-MIB for 6PE.

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

-- Revision history.
REVISION "201107060000Z" -- 06 July 2011 00:00:00 GMT
DESCRIPTION
    "Third published"

 ::= { 6peMplsStdMIB 1 }

-- 6PETunnelIfTable.

6PETunnelIfTable OBJECT-TYPE
SYNTAX      SEQUENCE OF TunnelIfEntry
MAX-ACCESS not-accessible
STATUS      current
DESCRIPTION
    "The (conceptual) table containing information on
    6PE tunnels."
 ::= { 6peMplsStdMIB 1 }

6PETunnelIfEntry OBJECT-TYPE
SYNTAX      TunnelIfEntry
MAX-ACCESS not-accessible
STATUS      current
DESCRIPTION
    "An entry (conceptual row) containing the information
    on a particular configured 6PE tunnel."
INDEX      { ifIndex }
 ::= { 6PETunnelIfTable 1 }

6PETunnelIfEntry ::= SEQUENCE {
    6PETunnelIfHopLimit      Integer32,
```

```
6PETunnelIfSecurity          INTEGER,
6PETunnelIfTOS               Integer32,
6PETunnelIfFlowLabel        IPv6FlowLabelOrAny,
6PETunnelIfLocalAddress     InetAddress,
6PETunnelIfRemoteAddress    InetAddress
}

6PETunnelIfLocalAddress OBJECT-TYPE
SYNTAX      IPAddress
MAX-ACCESS  read-only
STATUS      deprecated
DESCRIPTION
    "The address of the local endpoint of the tunnel"
 ::= { 6PETunnelIfEntry 1 }

6PETunnelIfRemoteAddress OBJECT-TYPE
SYNTAX      IPAddress
MAX-ACCESS  read-only
STATUS      deprecated
DESCRIPTION
    "The address of the remote endpoint of the tunnel"
 ::= { 6PETunnelIfEntry 2 }

6PETunnelIfHopLimit OBJECT-TYPE
SYNTAX      Integer32 (0 | 1..255)
MAX-ACCESS  read-write
STATUS      current
DESCRIPTION
    "The IPv6 Hop Limit to use in IPv6 header."
 ::= { 6PETunnelIfEntry 3 }

6PETunnelIfTOS OBJECT-TYPE
SYNTAX      Integer32 (-2..63)
MAX-ACCESS  read-write
STATUS      current
DESCRIPTION
    "The method used to set IPv6 Traffic Class in IP header."
 ::= { 6PETunnelIfEntry 4 }

        6PETunnelIfFlowLabel OBJECT-TYPE
SYNTAX      IPv6FlowLabelOrAny
MAX-ACCESS  read-write
STATUS      current
DESCRIPTION
    "The method used to set the IPv6 Flow Label value."
 ::= { 6PETunnelIfEntry 5 }
```

```
-- 6PEmplsIfTable.

6PEmplsIfTable    OBJECT-TYPE
SYNTAX            SEQUENCE OF PwMplsEntry
MAX-ACCESS        not-accessible
STATUS            current
DESCRIPTION
    "This table controls MPLS-specific parameters when the 6PE is
    going to be carried over MPLS cloud."
 ::= { 6peMplsStdMIB 2 }

6PEmplsEntry      OBJECT-TYPE
SYNTAX            6PEmplsEntry
MAX-ACCESS        not-accessible
STATUS            current
DESCRIPTION
    "A row in this table represents parameters specific to MPLS
    cloud for 6PE."

INDEX             { 6PEIndex }

 ::= { 6PEmplsIfTable 1 }
```

Figure 1

5. Security Considerations

It needs to be further identified.

6. IANA Considerations

This memo includes no request to IANA.

7. Normative References

- [RFC2578] McCloghrie, K., Ed., Perkins, D., Ed., and J. Schoenwaelder, Ed., "Structure of Management Information Version 2 (SMIv2)", STD 58, RFC 2578, April 1999.
- [RFC2579] McCloghrie, K., Ed., Perkins, D., Ed., and J. Schoenwaelder, Ed., "Textual Conventions for SMIv2", STD 58, RFC 2579, April 1999.
- [RFC2580] McCloghrie, K., Perkins, D., and J. Schoenwaelder,

"Conformance Statements for SMIV2", STD 58, RFC 2580, April 1999.

- [RFC2863] McCloghrie, K. and F. Kastenholz, "The Interfaces Group MIB", RFC 2863, June 2000.
- [RFC3410] Case, J., Mundy, R., Partain, D., and B. Stewart, "Introduction and Applicability Statements for Internet-Standard Management Framework", RFC 3410, December 2002.
- [RFC3811] Nadeau, T. and J. Cucchiara, "Definitions of Textual Conventions (TCs) for Multiprotocol Label Switching (MPLS) Management", RFC 3811, 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.
- [RFC4001] Daniele, M., Haberman, B., Routhier, S., and J. Schoenwaelder, "Textual Conventions for Internet Network Addresses", RFC 4001, February 2005.
- [RFC4087] Thaler, D., "IP Tunnel MIB", RFC 4087, June 2005.
- [RFC4382] Nadeau, T. and H. van der Linde, "MPLS/BGP Layer 3 Virtual Private Network (VPN) Management Information Base", RFC 4382, February 2006.
- [RFC4789] Schoenwaelder, J. and T. Jeffree, "Simple Network Management Protocol (SNMP) over IEEE 802 Networks", RFC 4789, November 2006.

Authors' Addresses

Gang Chen
China Mobile
53A, Xibianmennei Ave.,
Xuanwu District,
Beijing 100053
China

Email: chengang@chinamobile.com

Lianyuan Li
China Mobile
53A, Xibianmennei Ave.
Beijing 100053
P.R.China

Phone: +86-13910750201
Email: lilianyuan@chinamobile.com

Internet Engineering Task Force
Internet-Draft
Intended status: Standards Track
Expires: January 12, 2012

H. Chen
Huawei Technologies
N. So
Verizon Inc.
July 11, 2011

Extensions to RSVP-TE for P2MP LSP Egress Local Protection
draft-chen-mppls-p2mp-egress-protection-03.txt

Abstract

This document describes extensions to Resource Reservation Protocol - Traffic Engineering (RSVP-TE) for locally protecting egress nodes of a Traffic Engineered (TE) point-to-multipoint (P2MP) Label Switched Path (LSP) in a Multi-Protocol Label Switching (MPLS) and Generalized MPLS (GMPLS) network.

Status of this Memo

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

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

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

This Internet-Draft will expire on January 12, 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	3
2. Terminology	3
3. Conventions Used in This Document	3
4. Mechanism	3
4.1. An Example of Egress Local Protection	4
4.2. Set up of Backup sub LSP	5
4.3. Forwarding State for Backup sub LSP(s)	5
4.4. Detection of Egress Node Failure	5
5. Representation of a backup Sub LSP	6
5.1. EGRESS_BACKUP_SUB_LSP Object	6
5.1.1. EGRESS_BACKUP_SUB_LSP IPv4 Object	6
5.1.2. EGRESS_BACKUP_SUB_LSP IPv6 Object	7
5.2. EGRESS_BACKUP_SECONDARY_EXPLICIT_ROUTE Object	8
6. Path Message	8
6.1. Format of Path Message	8
6.2. Processing of Path Message	9
7. IANA Considerations	10
8. Acknowledgement	10
9. References	10
9.1. Normative References	10
9.2. Informative References	11
Authors' Addresses	11

1. Introduction

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

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

This document defines extensions to RSVP-TE for locally protecting an egress node of a Traffic Engineered (TE) point-to-multipoint (P2MP) Label Switched Path through using a backup P2MP sub LSP. The same extensions and mechanism can also be used to protect the egress node of a RSVP-TE P2P LSP.

2. Terminology

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

3. Conventions Used in This Document

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

4. Mechanism

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

4.1. An Example of Egress Local Protection

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

La, Lb and Lc are the designated backup egress/leaf nodes for the egress/leaf nodes L1, L2 and L3 of the P2MP LSP respectively. The backup sub LSP used to protect egress node L1 is from its previous hop node R3 to the backup node La. The backup sub LSP used to protect the egress node L2 is from its previous hop node R5 to the backup egress node Lb. The backup sub LSP used to protect the egress node L3 is from its previous hop node R5 to the backup egress node Lc via intermediate node Rc.

During normal operation, the traffic transported by the P2MP LSP is forwarded through R3 to L1, then delivered to its destination CE1. When the failure of L1 is detected, R3 forwards the traffic to the backup egress node La, which then delivers the traffic to its destination CE1. L1's failure CAN be detected by a BFD session between L1 and R3.

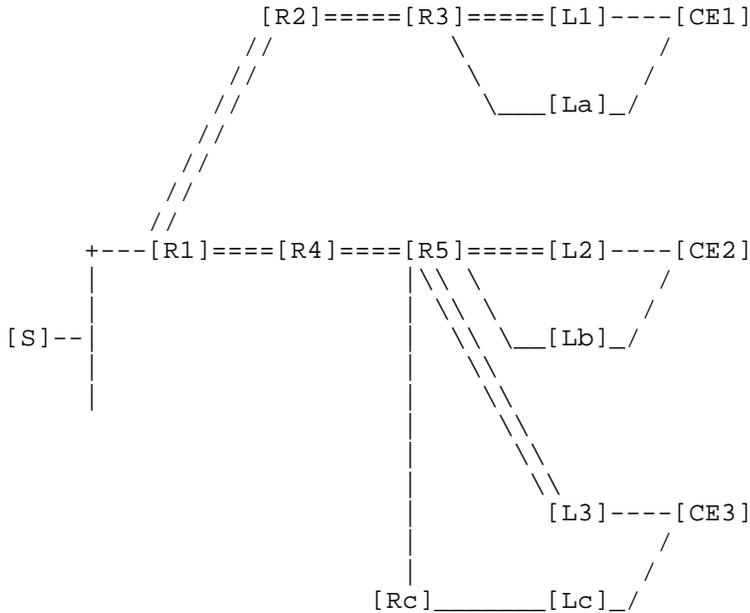


Figure 1: P2MP sub LSP for Locally Protecting Egress

4.2. Set up of Backup sub LSP

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

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

4.3. Forwarding State for Backup sub LSP(s)

The forwarding state for the backup sub LSP is different from that for a P2MP S2L sub LSP. After receiving the RSVP-TE RESV message for the backup sub LSP, the previous hop node creates a forwarding entry with an inactive state or flag called inactive forwarding entry. This inactive forwarding entry is not used to forward any data traffic during normal operations. It SHALL only be used after the failure of the protected egress node.

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

4.4. Detection of Egress Node Failure

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

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

Failure of the protected egress node and the link between itself and

its previous hop node CAN be detected through a BFD session between itself and its previous hop node.

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

Upon detecting any above mentioned failures, the previous hop node imports the traffic from the LSP into the backup sub LSP. The traffic is then delivered to its destination through the backup egress node.

5. Representation of a backup Sub LSP

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

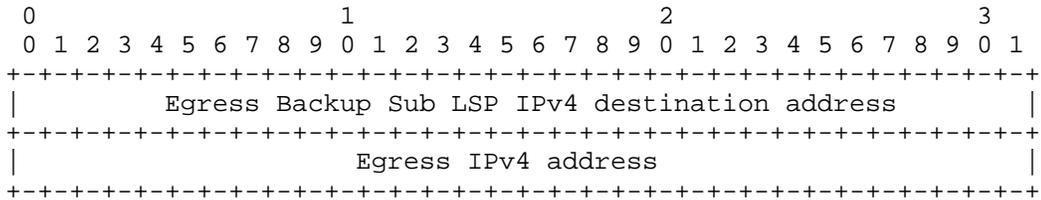
An EGRESS_BACKUP_SECONDARY_EXPLICIT_ROUTE Object (EB-SERO) is used to optionally specify the explicit route of a backup sub LSP that is from a previous hop node to a backup egress node. The EB-SERO is defined in the following section.

5.1. EGRESS_BACKUP_SUB_LSP Object

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

5.1.1. EGRESS_BACKUP_SUB_LSP IPv4 Object

EGRESS_BACKUP_SUB_LSP Class = 50,
EGRESS_BACKUP_SUB_LSP_IPv4 C-Type = 3

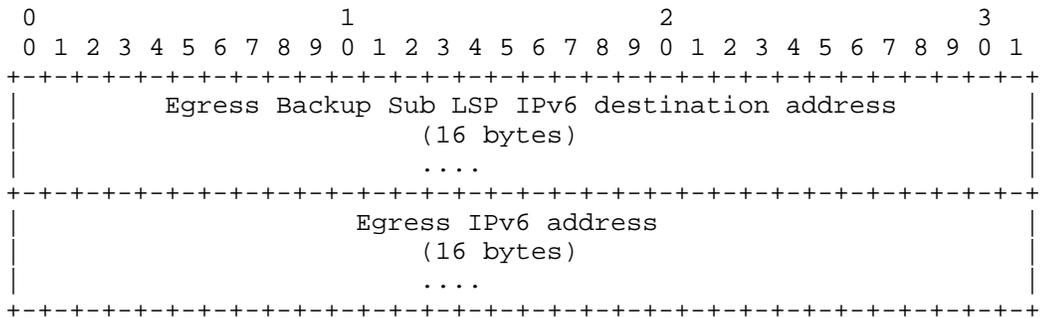


Egress Backup Sub LSP IPv4 destination address
IPv4 address of the backup sub LSP destination is the backup egress node.
Egress IPv4 address
IPv4 address of the egress node

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

5.1.1.2. EGRESS_BACKUP_SUB_LSP IPv6 Object

EGRESS_BACKUP_SUB_LSP Class = 50,
EGRESS_BACKUP_SUB_LSP_IPv6 C-Type = 4



Egress Backup Sub LSP IPv6 destination address
IPv6 address of the backup sub LSP destination is the backup egress node.
Egress IPv6 address
IPv6 address of the egress node

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

5.2. EGRESS_BACKUP_SECONDARY_EXPLICIT_ROUTE Object

The format of an EGRESS_BACKUP_SECONDARY_EXPLICIT_ROUTE (EB-SERO) object is defined as identical to that of the ERO. The class of the EB-SERO is the same as the SERO defined in RFC 4873. The EB-SERO uses a new C-Type = 3, or may use another number assigned by Internet Assigned Numbers Authority (IANA). The formats of sub-objects in an EB-SERO are identical to those of sub-objects in an ERO defined in RFC 3209.

6. Path Message

This section describes extensions to the Path message defined in RFC 4875. The Path message is enhanced to transport the information about a backup egress node to the previous hop node of an egress node of a P2MP LSP through including an egress backup sub LSP descriptor list.

6.1. Format of Path Message

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 sub LSP descriptor list>]

```

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

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

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

6.2. Processing of Path Message

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

An intermediate node (a transit or branch node) receives the Path message with an egress backup sub LSP descriptor list. Then it MUST put the EGRESS_BACKUP_SUB_LSP (according to EB-SERO if exists) into the Path message. This SHALL be done for each EGRESS_BACKUP_SUB_LSP containing a backup egress node in the list. After that, the message is sent to the previous hop node of the protected egress node. If the intermediate node is the previous hop node of the protected egress node, it generates a new Path message based on the information in the EGRESS_BACKUP_SUB_LSP (and according to EB-SERO if exists) upon receiving the Path message with the EGRESS_BACKUP_SUB_LSP containing the backup egress node.

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

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

7. IANA Considerations

TBD

8. Acknowledgement

The author would like to thank Richard Li and Quintin Zhao for their valuable comments on this draft.

9. References

9.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC3692] Narten, T., "Assigning Experimental and Testing Numbers Considered Useful", BCP 82, RFC 3692, January 2004.
- [RFC2205] Braden, B., Zhang, L., Berson, S., Herzog, S., and S. Jamin, "Resource ReSerVation Protocol (RSVP) -- Version 1 Functional Specification", RFC 2205, September 1997.
- [RFC3031] Rosen, E., Viswanathan, A., and R. Callon, "Multiprotocol Label Switching Architecture", RFC 3031, January 2001.
- [RFC3209] Awduche, D., Berger, L., Gan, D., Li, T., Srinivasan, V., and G. Swallow, "RSVP-TE: Extensions to RSVP for LSP Tunnels", RFC 3209, December 2001.
- [RFC3473] Berger, L., "Generalized Multi-Protocol Label Switching (GMPLS) Signaling Resource ReserVation Protocol-Traffic Engineering (RSVP-TE) Extensions", RFC 3473, January 2003.
- [RFC4090] Pan, P., Swallow, G., and A. Atlas, "Fast Reroute Extensions to RSVP-TE for LSP Tunnels", RFC 4090, May 2005.
- [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.

9.2. Informative References

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

Authors' Addresses

Huaimo Chen
Huawei Technologies
Boston, MA
USA

Email: Huaimochen@huawei.com

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

Email: Ning.So@verizonbusiness.com

Internet Engineering Task Force
Internet-Draft
Intended status: Standards Track
Expires: January 12, 2012

H. Chen
Huawei Technologies
N. So
Verizon Inc.
July 11, 2011

Extensions to RSVP-TE for P2MP LSP Ingress Local Protection
draft-chen-mppls-p2mp-ingress-protection-03.txt

Abstract

This document describes extensions to Resource Reservation Protocol - Traffic Engineering (RSVP-TE) for locally protecting the ingress node of a Traffic Engineered (TE) Point-to-MultiPoint (P2MP) Label Switched Path (LSP) in a Multi-Protocol Label Switching (MPLS) and Generalized MPLS (GMPLS) network.

Status of this Memo

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

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

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

This Internet-Draft will expire on January 12, 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	3
2. Terminology	4
3. Conventions Used in This Document	4
4. Mechanism	4
4.1. An Example of Ingress Local Protection	4
4.2. Set up of Backup P2MP sub Tree	5
4.3. Forwarding State for Backup P2MP sub Tree	5
4.4. Detection of Failure around Ingress	6
5. LSP Information Message	6
5.1. Format of LSP Information Message	7
5.2. Processing of LSP Information Message	7
6. LSP Information Confirmation Message	8
6.1. Format of LSP Information Confirmation Message	8
6.2. Processing of LSP Information Confirmation Message	8
7. PATH Messages for Backup P2MP sub Tree	9
7.1. Construction of PATH Messages	9
7.2. Processing of PATH Messages	9
8. IANA Considerations	10
9. Acknowledgement	10
10. References	10
10.1. Normative References	10
10.2. Informative References	11
Authors' Addresses	11

1. Introduction

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

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

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

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

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

2. Terminology

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

3. Conventions Used in This Document

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

4. Mechanism

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

4.1. An Example of Ingress Local Protection

Figure 1 below illustrates an example of using a backup P2MP sub tree to locally protect the ingress of a P2MP LSP. The P2MP LSP to be protected is from ingress node R1 to three egress/leaf nodes: L1, L2 and L3. The backup P2MP sub tree used to protect the ingress node R1 is from backup ingress node Ra to the next hop nodes R2 and R4 of the ingress node R1 along the P2MP LSP. The traffic from source S may be delivered to both R1 and Ra. R1 introduces the traffic into the P2MP LSP, which is sent to the egress/leaf nodes L1, L2 and L3 along the P2MP LSP. Ra normally does not put the traffic into the backup P2MP sub tree, which is from Ra to R2 and R4. There may be a BFD session between ingress node R1 and backup ingress node Ra. Ra uses this BFD session to detect the failure of ingress R1. When Ra detects the failure of R1, it imports the traffic from the source S into the backup P2MP sub tree. The traffic from the sub tree is merged into the P2MP LSP at R2 and R4, and then sent to the egress/leaf nodes L1, L2 and L3 along the P2MP LSP.

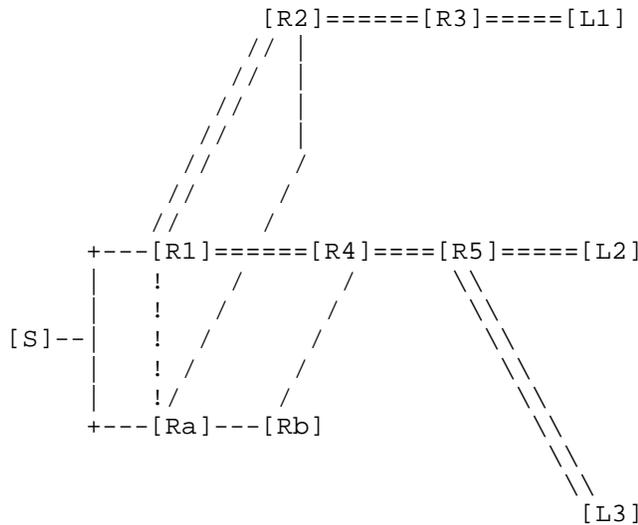


Figure 1: P2MP sub Tree for Locally Protecting Ingress

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

4.2. Set up of Backup P2MP sub Tree

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

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

4.3. Forwarding State for Backup P2MP sub Tree

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

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

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

4.4. Detection of Failure around Ingress

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

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

A failure of the ingress node can be detected through a BFD session between the ingress node and the backup ingress node. A failure of the link between the source node and the ingress node can be detected by a BFD session running on the link.

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

5. LSP Information Message

LSP information messages are used to transfer the information about a P2MP LSP to a backup ingress node from an ingress node. This section describes the format of an LSP information message and processing of the message.

5.1. Format of LSP Information Message

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

```

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

```

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

The value of the Msg Type field in the common header in the P2MP LSP information message will be a new number such as 68 for the LSP information message, or may be another number assigned by Internet Assigned Numbers Authority (IANA).

5.2. Processing of LSP Information Message

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

When the backup ingress node receives the RSVP-TE LSP information message from the ingress node, the backup ingress node stores the LSP information, constructs PATH messages, and sends the PATH messages downstream accordingly. If the backup ingress node has not received any RSVP-TE LSP information message for an extended period of time

(e.g. a cleanup timeout interval), the backup ingress node SHALL remove the information about the P2MP LSP, constructs PathTear messages, and send the PathTear messages downstream accordingly.

6. LSP Information Confirmation Message

LSP information confirmation messages are used to confirm that the corresponding LSP information messages are received. With the confirmation messages, the refresh of the LSP information messages is not needed. This section describes the format of an LSP information confirmation message and processing of the message.

6.1. Format of LSP Information Confirmation Message

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

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

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

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

6.2. Processing of LSP Information Confirmation Message

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

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

7. PATH Messages for Backup P2MP sub Tree

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

7.1. Construction of PATH Messages

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

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

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

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

The construction of a PATH message on a backup ingress node for a backup P2MP sub tree is similar to the construction of a normal PATH message on an ingress node for a P2MP LSP. It is based on LSP information messages and a computed path for the backup P2MP sub tree.

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

7.2. Processing of PATH Messages

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

processing of PATH messages for a P2MP LSP.

8. IANA Considerations

TBD

9. Acknowledgement

The author would like to thank Richard Li and Quintin Zhao for their valuable comments on this draft.

10. References

10.1. Normative References

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

10.2. Informative References

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

Authors' Addresses

Huaimo Chen
Huawei Technologies
Boston, MA
USA

Email: Huaimochen@huawei.com

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

Email: Ning.So@verizonbusiness.com

MPLS Working Group
Internet Draft
Intended status: Standards Track
Created: April 25, 2011
Expires: October 25, 2011

Tae-sik Cheung
Jeong-dong Ryoo
ETRI

MPLS-TP Shared Mesh Protection
draft-cheung-mpls-tp-mesh-protection-03.txt

Abstract

This document describes a mechanism to address the requirement for protection of Label Switched Paths (LSPs) in an MPLS Transport Profile (MPLS-TP) mesh topology. The shared mesh protection mechanism enables multiple protection paths within a shared mesh protection domain to share protection resources for the protection of working paths by coordinating protection switching operations according to the priority assigned to each end-to-end linear protection domain.

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

Status of this Memo

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

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

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

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

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

This Internet-Draft will expire on October 25, 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.....	3
2. Conventions used in this document.....	4
2.1. Acronyms.....	4
2.2. Definitions and Terminology.....	5
3. Shared Mesh Protection.....	5
3.1. Protection Switching Priority.....	6
3.2. Shared Start and End Nodes.....	6
3.3. Bridge and Selector Models.....	8
3.4. Shared Mesh Protection Planning.....	9
3.5. Shared Mesh Protection Switching.....	9
3.5.1. Protection Switching Event.....	10
3.5.2. Protection Locking.....	11
4. Protocol.....	11
4.1. PDU Format.....	11
4.1.1. Protection Switching Event Message.....	12
4.1.2. Protection Locking Message.....	12
4.2. Message Transmission.....	13
5. Operation of Shared Mesh Protection.....	13
6. Manageability Considerations.....	16
7. Security Considerations.....	16
8. IANA Considerations.....	16
9. References.....	16
9.1. Normative References.....	16
9.2. Informative References.....	16

1. Introduction

The MPLS Transport Profile (MPLS-TP) is a packet transport technology based on a profile of the MPLS and Pseudowires (PW) [RFC3031], [RFC3985], and [RFC5085]. MPLS-TP is the application of MPLS to the construction of packet-switched paths that are analogous to traditional circuit-switched technologies. Requirements for MPLS-TP are specified in [RFC5654].

An important feature of a transport network is its survivability function and the ability to maintain or recover traffic following a network failure or attack. According to Requirement 56 of [RFC5654], MPLS-TP must provide protection and restoration mechanisms, and it must also be possible to require protection of 100% of the traffic on the protected path (Requirement 58).

1+1 and 1:1 protection can meet these requirements by reserving the equivalent amount of network resources for the protection paths as is used by the normal traffic to be protected. While those dedicated protection mechanisms provide very good protection capabilities, they are resource inefficient and will increase overall network resource consumption. Deploying 1+1 and 1:1 protection mechanisms for all services that require resiliency, dramatically increases network costs.

[RFC5654] also establishes that MPLS-TP should support shared protection (Requirement 68). 1:n end-to-end protection uses one protection path to protect n working paths. This improves overall network utilization, but the resource (bandwidth) allocated to a protection path is typically not sufficient to protect multiple and simultaneous failures on different working paths. If multiple working paths are required to be switched to a protection path concurrently, the path with the highest priority should be protected first as described in [I-D.ietf-mpls-tp-survive-fwk].

In 1+1 and 1:1 protection, the end nodes of the working path must be the same as those of the protection path. The same applies in 1:n protection where all pairs of end nodes of the n working paths are the same, and the protection path must also have the same end nodes. In the event that the MPLS-TP network scales up, the number of Label Switched Paths (LSPs) having different end nodes will also increase. The network utilization benefit for sharing protection resources among multiple protected domains for such LSPs will increase accordingly.

Requirement 68 of [RFC5654] specifies that MPLS-TP should support 1:n shared mesh recovery, and Requirement 69 states that MPLS-TP must support sharing of protection resources. It may be possible that some working paths are sufficiently disjoint and would be unlikely to be simultaneously affected by a single network failure. Typically, such a scenario is hard to track in real network environments where new services are often added and removed.

In mesh protection, network resources may be shared to provide protection for working paths that do not share the same end nodes at the edge of a protection domain. This form of protection can make very efficient use of network resources, but requires careful synchronization to ensure that only one set of traffic is switched to the protection resources at any one time.

[RFC4428] defines two shared mesh recovery schemes named $(1:1)^n$ and $(M:N)^n$. $(1:1)^n$ recovery scheme is a simple case of $(M:N)^n$ recovery scheme. In $(1:1)^n$ protection, n working paths are protected by n dedicated protection paths while sharing the same protection bandwidth.

The protection bandwidth can be optimized to allow only one of the n working paths to be protected at the same time. In this case, it achieves same amount of network utilization with $1:n$ protection.

$(1:1)^n$ protection defined in [RFC4428] differs with that defined in [G.808.1] in that the former allows each n pairs of working and protection paths to have different end nodes while the latter applies to the case where all pairs have same end nodes.

This document defines a shared mesh protection mechanism based on the concept of $(1:1)^n$ recovery scheme defined in [RFC4428] and a coordination to share protection resource based on the protection switching priority assigned to each pair of working and protection paths. Each working path is protected by its own end-to-end linear protection protocol.

The shared mesh protection mechanism defined in this document utilizes any existing MPLS-TP linear protection switching mechanisms being developed in the context of MPLS-TP, and assumes that the protection paths are established and ready to forward data prior to a failure. Upon detection of a failure on a working path, only two end nodes of the failed working path exchange their linear protection protocol messages to switch data traffic. No explicit activation procedure to switch data traffic to the protection path is needed in the intermediate nodes along the protection path.

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].

2.1. Acronyms

This document uses the following acronyms:

LoP	Lockout of Protection
LSP	Label Switched Path

MIP	Maintenance Entity Group Intermediate Point
MPLS-TP	MPLS Transport Profile
P2MP	Point-to-multipoint
P2P	Point-to-point
PW	Pseudowire
SEN	Shared End Node
SSN	Shared Start Node

2.2. Definitions and Terminology

This document defines two protection domains as follows:

- o End-to-end linear protection domain: A protection domain as defined in [I-D.ietf-mpls-tp-survive-fwk] for protecting a P2P or P2MP LSP. It consists of two or more end nodes at the boundary of the domain and a working path and a number of protection paths between the end nodes. An end-to-end linear protection switching protocol runs within the domain.
- o Shared mesh protection domain: A protection domain for protecting a number of P2P or P2MP LSPs. It consists of a number of end-to-end linear protection domains. Each end-to-end linear protection domain shares protection resources with other domains. The shared protection resource may be a node, link, transport path segment or concatenated transport path segment. A shared mesh protection switching protocol runs within the domain.

3. Shared Mesh Protection

Figure 1 shows a simple case of shared mesh protection. Consider two paths ABCDE and VWXYZ. These paths do not share end points so they cannot make use of 1:n protection even though they also do not share any common points of failure.

ABCDE may be protected by the path APQRE, and VWXYZ can be protected by the path VPQRZ. For both cases, 1:1 protection may be used. If there are no failures affecting either of the two working paths, the network segment PQR carries no traffic. In the event of only one failure, the segment PQR carries traffic from the working path that experiences the failure.

Thus, it is possible for the network resources on the segment PQR to be shared by the two protection paths. In this way, shared mesh protection can substantially reduce the amount of network resources that have to be reserved to provide protection of a 1:n nature.

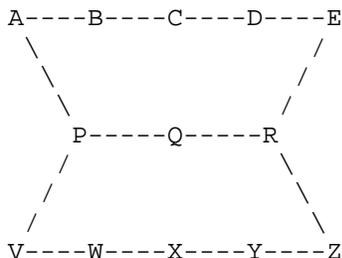


Figure 1 : A Shared Mesh Protection Topology

The shared mesh protection domain shown in Figure 1 has two end-to-end linear protection domains. One consists of two end nodes A and E and one working path ABCDE and one dedicated protection path APQRE. And the other consists of end nodes V and Z and one working path VWXYZ and one dedicated protection path VPQRZ. Those two domains share a protection segment PQR.

3.1. Protection Switching Priority

A ?Protection Switching Priority? needs to be defined for each end-to-end linear protection domain that has a protection path sharing the same protection resource. According to the Protection Switching Priority, a protection path can displace the other protection path already using the shared protection resources and protect its own working path.

The Protection Switching Priority may be provisioned by the network management system. By default, equal priority is assumed resulting in first-come first-served recovery. If multiple end-to-end linear protection domains request protection switching simultaneously, a pre-defined identifier **MUST** be used to give priority among them. The definition of the identifier is for further study.

3.2. Shared Start and End Nodes

A Shared Start Node (SSN) is the first node of a unidirectional shared protection segment. For example, in Figure 1, node P is a SSN on unidirectional protection paths A->P->Q->R->E and V->P->Q->R->Z. In this version of document, SSN does not involve in the shared mesh protection operation. SSN may act as a Maintenance Intermediate Point (MIP) for each protection path sharing the same protection resources.

Similarly, a Shared End Node (SEN) is defined as the last node of a unidirectional shared protection segment (for example, node R on unidirectional protection paths A->P->Q->R->E and V->P->Q->R->Z in Figure 1). SEN involves in the shared mesh protection operation for coordinating the use of the unidirectional shared protection segment. A SEN acts as a MIP on each protection path that shares the protection resource.

Table 1 summarizes the relationship between SSN and SEN of the shared protection segment and protection paths sharing it.

Table 1: SSN/SEN in Figure 1

Protection paths	Shared protection segment	SSN	SEN
A->P->Q->R->E, V->P->Q->R->Z	P->Q->R	P	R
E->R->Q->P->A, Z->R->Q->P->V	R->Q->P	R	P

Figure 2 shows a more complex example of the shared mesh protection domain. Three working paths ABC, DEF, and GHJ are protected by the protection paths APQC, DRSF, and GPQRSJ, respectively.

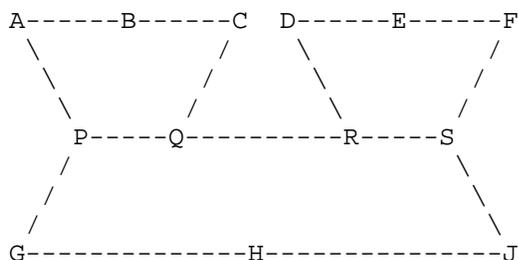


Figure 2: A More Complex Mesh Protection Example

In this example, the unidirectional protection path G->P->Q->R->S->J shares resources with two other protection paths, and both P and R are SSNs, while Q and S are SENs. (See Table 2.)

Table 2: SSN/SEN in Figure 2

Protection paths	Shared protection segment	SSN	SEN
A->P->Q->C, G->P->Q->R->S->J	P->Q	P	Q
C->Q->P->A, J->S->R->Q->P->G	Q->P	Q	P
D->R->S->F, G->P->Q->R->S->J	R->S	R	S
F->S->R->D, J->S->R->Q->P->G	S->R	S	R

3.3. Bridge and Selector Models

Figure 3 shows bridge and selector model for nodes in the shared mesh protection topology shown in Figure 1. For simplicity, only end nodes and shared nodes are modelled. Figure 3 illustrates that node A and E send and receive normal traffic (1) through protection path (1) and node V and Z send and receive normal traffic (2) through working path (2).

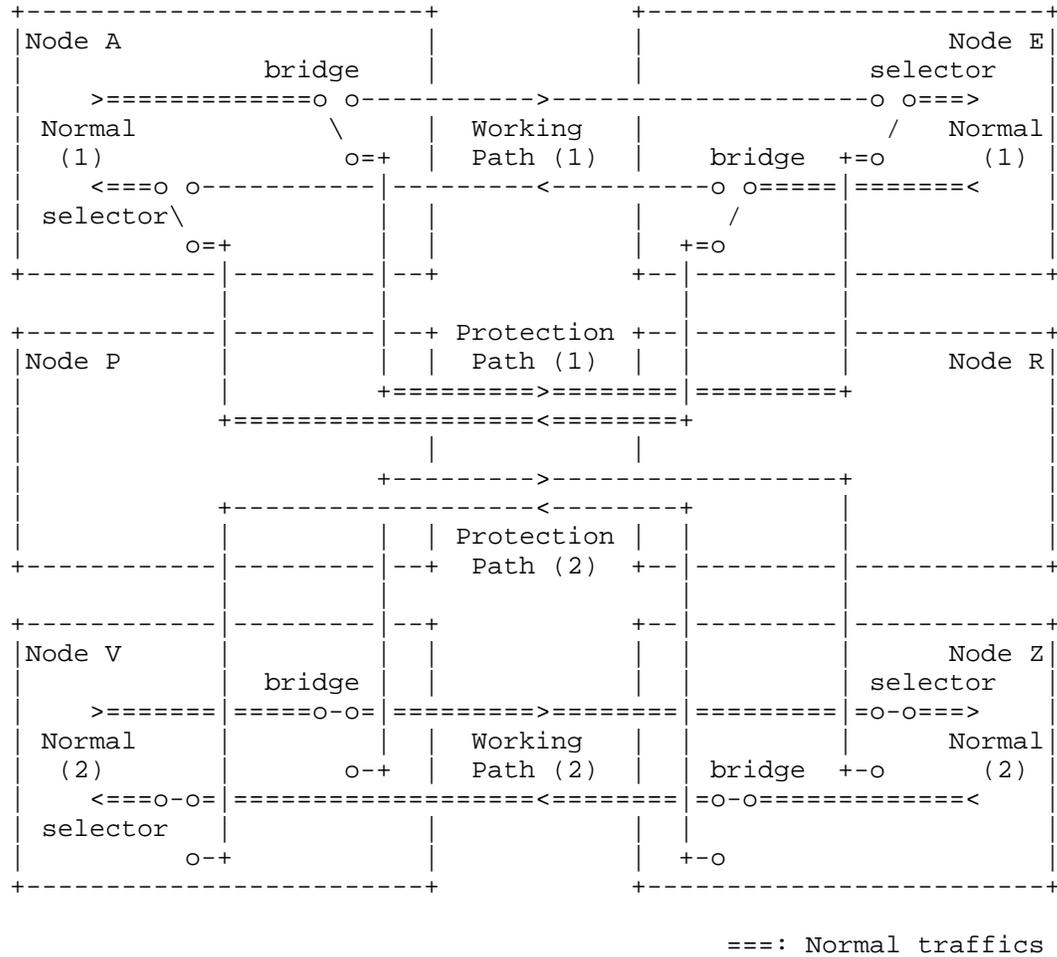


Figure 3: Bridge and selector model of the example mesh protection topology shown in Figure 1

Each end node has a bridge and selector to send and receive normal traffic through its working or protection path. Shared nodes have no bridge or selector and all the protection paths are pre-provisioned and monitored.

The bandwidth occupied by each working path is $B_{Wi} = B_{Ni} + B_{OAM_{Wi}}$; i.e., the bandwidth for the normal traffic signal #i, plus the bandwidth for the OAM used to monitor the working path #i. The bandwidth of the shared protection segment required to protect at least one normal traffic signal among those flowing through n working paths is calculated as:

$$B_P = \text{MAX}(B_{N1}, B_{N2}, \dots, B_{Nn}) + (B_{OAM_{P1}} + B_{OAM_{P2}} + \dots + B_{OAM_{Pn}}).$$

The bridge and selector model of shared mesh protection is similar to that for (1:1)ⁿ protection type defined in [G.808.1], but it differs in that each working path connects different pair of end nodes, and each protection path shares a same protection segment.

3.4. Shared Mesh Protection Planning

Shared mesh protection will typically be subject to careful network planning. This will include:

- o Determining which working paths are disjoint and so will not be subject to common failures.
- o Assigning Protection Switching Priority to each end-to-end linear protection domain so that the protection paths can be activated correctly.
- o Ensuring that working paths of high Protection Switching Priority do not share resources on their protection paths in such a way that would mean that one of them could not be protected.
- o Enabling the necessary shared mesh protection functions at SEN.

Note that some control plane features of GMPLS may be used to dynamically install shared mesh protection. These features are out of scope for this document which focuses on the operation of shared mesh protection switching once it has been installed.

3.5. Shared Mesh Protection Switching

The shared mesh protection mechanism is designed to fully utilize the existing end-to-end linear protection switching without any changes except the following two additional functionalities:

- o Function to generate a protection switching event message to the SEN when a switching action occurs at the end-to-end linear protection domain.
- o Function to take a protection locking message from the SEN, and incorporate it as the Lockout of Protection (LoP) command.

3.5.1. Protection Switching Event

If an end node of a working path detects a failure condition, it triggers the protection switching and exchanges linear protection switching protocol messages with its peer end node at the other end of the working/protection path according to the operation defined in its own linear protection switching, which is independent of the mesh protection switching mechanism specified in this document.

At the same time, for the shared mesh protection, the end node notifies its protection switching event to SENSs by sending a protection switching event message.

The protection switching event message MUST be transmitted immediately when an end node changes its selector position either from working to protection or vice versa.

If an end-to-end linear protection domain operates in a bidirectional protection switching, both end nodes will change their bridge and selector positions even when a unidirectional failure is detected on one end node, and therefore, both end nodes will transmit the messages to their corresponding SENSs.

If an end-to-end linear protection domain operates in a unidirectional protection switching and a unidirectional failure is detected, the end node that detects the failure will change its selector position and send the messages to its corresponding SENSs.

There are two possible ways that the protection switching event message could be delivered to SENSs.

- o Option 1 (Default): Use a P2P message

The end node of the protection path that is becoming active sends messages directly to each SEN. The path from an end node to a SEN is a segment of the protection path and the messages are delivered to each SEN by properly setting the TTL values of the messages for each SEN. This ensures fate sharing of the messages with other OAM or data traffics. Alternatively, an end node may have dedicated paths to communicate with each SEN. The option 1 requires N messages to be sent if N SENSs exist on the protection path. Furthermore, it requires that the end nodes of the protection path know about all SENSs - this is perfectly possible in simple configurations or through the use of a dynamic control plane.

- o Option 2: Use a P2MP message

The end node sends a message similar to a route trace to the peer end node. It will be passed to all SENs. When a SEN receives the message, it will simultaneously take a copy of the message for local use, and forward a copy to the next hop.

An on-demand OAM message like route trace may contain the required information or the message itself may be transferred using a pre-provisioned P2MP LSP. In this option, the end node becomes a root node and all SENs and the peer end node become leaf nodes.

3.5.2. Protection Locking

If a SEN receives the protection switching event notifying that a protection switching has begun in an end-to-end linear protection domain, it compares the Protection Switching Priority of the protection domain notifying the event with those of other protection domains sharing the same protection segment.

The SEN does not take an action to the protection domains having higher priorities, but for those having equal or lower priorities, it sends protection locking messages to those end nodes to prevent any protection switching to be occurred.

When an end node receives the protection locking message from SEN, it will take the message as an input to the end-to-end linear protection switching, and follows the linear protection switching procedure to process end-to-end LoP command. Since the LoP command has the highest priority in the linear protection switching protocol, it will prohibit any further protection switching in the protection domain. If a protection domain having lower priority currently uses the shared protection segment, it will stop occupying the protection bandwidth by the command.

When a SEN receives a protection switching event message notifying the clearance of protection state from an end node, it sends a protection locking message to the end node to clear the LoP command.

4. Protocol

4.1. PDU Format

The shared mesh protection protocol messages MUST be sent over a G-ACh as defined in [RFC5586].

The shared mesh protection protocol messages are as follows:

- o Protection switching event message and
- o Protection locking message.

The channel type in ACH is used to indicate shared mesh protection protocol. The shared mesh protection protocol does not use ACH TLVs, therefore the protocol message MUST follow the ACH.

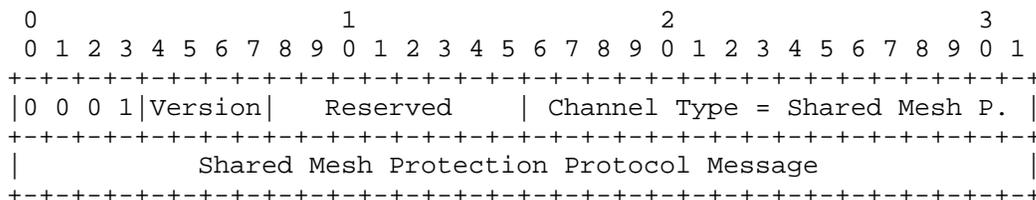


Figure 4: Shared mesh protection protocol message header

4.1.1. Protection Switching Event Message

The protection switching event message format is as follows:

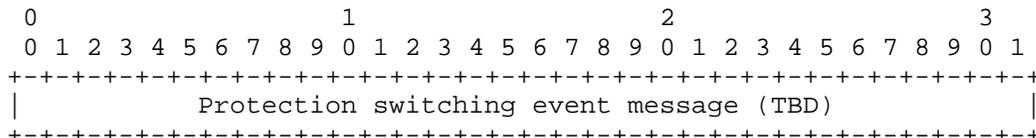


Figure 5: Protection switching event message format

In the message, following field will be provided:

- o Version
- o An Identifier of the end-to-end linear protection domain to which the end node sending this message belongs
- o Request/State identifying:
 - protection path is occupied by normal traffic, or
 - protection path is not occupied.

4.1.2. Protection Locking Message

The protection locking message format is as follows:

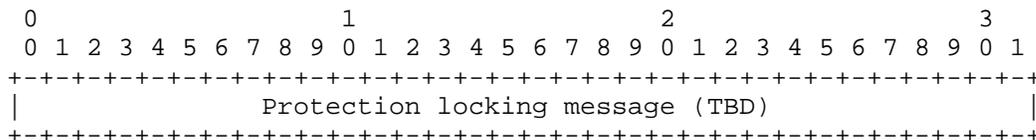


Figure 6: Protection locking message format

In the message, following field will be provided:

- o Version
- o An identifier of the end-to-end linear protection domain to which the end node receiving this message belongs
- o Request/State identifying:
 - protection lock requested or
 - protection unlock requested.

4.2. Message Transmission

A new message must be transmitted immediately. The first three messages should be transmitted as fast as possible so that fast protection switching is possible even if one or two messages are lost or corrupted. The interval of the first three messages should be less than 3.3ms. Messages after the first three should be transmitted with the interval of 5 seconds.

If no valid message is received, the last valid received information remains applicable.

5. Operation of Shared Mesh Protection

This section illustrates the operation of the shared mesh protection protocol for the exemplary topology shown in Figure 2 with following assumptions:

- o The shared mesh protection domain consists of following end-to-end linear protection domains (LPDs):
 - LPD1: Working path ABC (W1) / Protection path APQC (P1)
 - LPD2: Working path GHJ (W2) / Protection path GPQRSJ (P2)
 - LPD3: Working path DEF (W3) / Protection path DRSF (P3)
- o Protection Switching Priority is LPD1 > LPD2 > LPD3. (LPD1 has the highest priority.)
- o All working paths are protected by 1:1 bidirectional protection switching.

If a unidirectional failure occurs on the W2 in the direction from node H to node G as shown in Figure 7, the shared mesh protection will operate as follows:

- a. Node G detects the failure, and initiates the linear protection switching for the failed W2.

- b. At the same time, node G generates the protection switching event message saying that a protection switching event happened to node P and R, which are SENs for J->H->G.
- c. The SEN P compares the protection switching priority of LPD2 with that of LPD1. In this example, as the priority of LPD1 is higher than LPD2, SEN P does not take an action to node A.
The SEN R compares the protection switching priority of LPD2 with that of LPD3. In this example, as the priority of LPD3 is lower than LPD2, SEN R sends the protection locking message requesting LoP to node D.
- d. Node D takes the protection locking message as an input to the linear protection switching, and follows the linear protection switching procedure to process the end-to-end LoP command.

As LPD2 operates in a 1:1 bidirectional protection switching, node J also changes its bridge and selector state to synchronize with node G, thus it will generate the protection switching event message to node S and Q, which are SENs for G->H->J. By the same procedure described above, the SEN S sends the protection locking message to node F while the SEN Q does not take an action to node C.

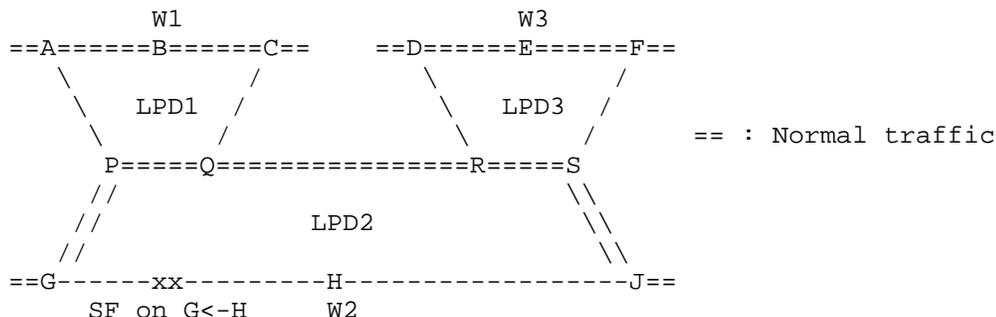


Figure 7 : Shared Mesh Protection Example 1

Figure 8 shows a progression from Figure 7. While LPD2 is in protecting state with its traffic following the protection path P2 (GPQRSJ), another unidirectional failure occurs on the W1 in the direction from node B to node A.

In this case, the shared mesh protection will operate as follows:

- a. Node A detects the failure, and initiates the linear protection switching for the failed W1.
- b. At the same time, node A generates the protection switching event message saying that a protection switching event happened to node P, which is SEN for C->B->A.

- c. The SEN P compares the protection switching priority of LPD1 with that of LPD2. In this example, as the priority of LPD2 is lower than LPD1, SEN P sends the protection locking message requesting LoP to node G.
- d. Node G takes the protection locking message as an input to the linear protection switching, and follows the linear protection switching procedure to process the end-to-end LoP command. When LPD2 is forced to lock its protection path P2, it may try to find another available path. m:n protection or other recovery mechanism can be used for this, but this discussion is out of scope of this document.
- e. As the node G changes its bridge and selector states from protection to working, it will generate the protection switching event message saying that a protection switching event has been cleared to node P and R, which are SENs for J->H->G.
- f. The SEN P compares the protection switching priority of LPD2 with that of LPD1 and does not take an action to node A, but the SEN R sends the protection locking message requesting clearance of LoP to node D.
- g. Node D takes the message as an input to the linear protection switching, and follows the linear protection switching procedure to clear the end-to-end LoP command.

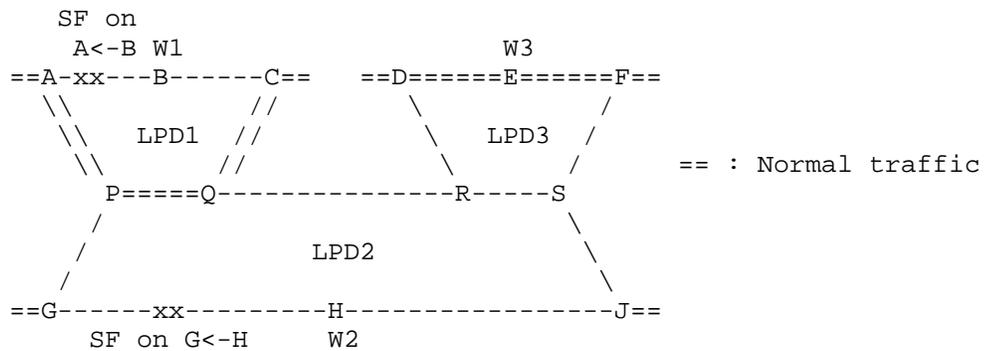


Figure 8 : Share Mesh Protection Example 2

6. Manageability Considerations

To be added in future version.

7. Security Considerations

To be added in future version.

8. IANA Considerations

To be added in future version.

9. References

9.1. Normative References

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

[RFC5654] Brungard, D., Betts, M., Sprecher, N. and Ueno, S., "Requirements of an MPLS Transport Profile", RFC5654, September 2009.

9.2. Informative References

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

[RFC3985] Bryant, S. and P. Pate, "Pseudo Wire Emulation Edge-to-Edge (PWE3) Architecture", RFC3985, March 2005.

[RFC5085] Nadeau, T. and Pignataro, C., "Pseudo Wire (PW) Virtual Circuit Connectivity Verification ((VCCV): A Control Channel for Pseudowires", RFC5085, December 2007.

[I-D.ietf-mpls-tp-survive-fwk] Sprecher, N. and Farrel A., "Multiprotocol Label Switching Transport Profile Survivability Framework", draft-ietf-mpls-tp-survive-fwk, work on progress.

[G.808.1] ITU-T, "Generic Protection Switching - Linear trail and subnetwork protection", Recommendation G.808.1, February 2010.

[RFC4428] Papadimitriou, D. and E. Mannie, "Analysis of Generalized Multi-Protocol Label Switching (GMPLS) ? based Recovery Mechanisms (including Protection and Restoration) Recovery (Protection and Restoration)", RFC 4428, March 2006.

Authors' Addresses

Tae-sik Cheung
ETRI
161 Gajeong, Yuseong, Daejeon, 305-700, South Korea

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

Jeong-dong Ryoo
ETRI
161 Gajeong, Yuseong, Daejeon, 305-700, South Korea

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

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: January 1, 2012

E. Osborne
Cisco
F. Zhang
ZTE
Y. Weingarten
Nokia Siemens Networks
June 30, 2011

MPLS-TP ltoN Protection
draft-ezy-mpls-lton-protection-00.txt

Abstract

As part of the Transport Profile for Multiprotocol Label Switching (MPLS-TP) there is a requirement to support 1:n linear protection for transport paths. This requirement is elaborated on in the MPLS-TP Survivability Framework document [SurvivFwk]. The basic protocol for linear protection was specified in the MPLS-TP Linear Protection document [LinProt] but is limited to 1+1 and 1:1 protection. This document extends the protocol defined there to address the additional functionality necessary to support scenarios of a single protection path preconfigured to provide protection of multiple transport paths between two joint endpoints.

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

Status of this Memo

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

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

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

This Internet-Draft will expire on January 1, 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.

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

Table of Contents

1.	Introduction	4
1.1.	1:n Protection architecture	4
1.2.	Path priority	5
1.3.	Preemption	6
1.4.	Contributing authors	6
2.	Conventions used in this document	6
2.1.	Acronyms	7
2.2.	Definitions and Terminology	7
3.	Changes to PSC	7
3.1.	PSC	8
3.2.	Changes to PSC Payload	8
3.2.1.	Acknowledge (K) flag	9
3.2.2.	Fault path (FPath) field	9
3.2.3.	Data path (Path) field	9
3.3.	Changes to PSC Operation	10
3.3.1.	Basic operation	10
3.3.2.	Two-phased operation	10
3.3.3.	Acknowledge message	11
3.3.4.	Wait for Acknowledge (Wfa) timer	12
3.3.5.	Additional PSC State	12
4.	IANA Considerations	15
5.	Security Considerations	15
6.	Acknowledgements	16
7.	References	16
7.1.	Normative References	16
7.2.	Informative References	16
	Appendix A. PSC state machine tables	17
	Authors' Addresses	20

1. Introduction

The MPLS Transport Profile (MPLS-TP) Requirements document [TPReq] includes requirements for the necessary survivability tools that are required for MPLS based transport networks. Network survivability is the ability of a network to recover traffic delivery following failure, or degradation of network resources. Requirement 67 lists various types of 1:n protection architectures that are required for MPLS-TP. The MPLS-TP Survivability Framework [SurvivFwk] is a framework for survivability in MPLS-TP networks, and describes recovery elements, types, methods, and topological considerations, focusing on mechanisms for recovering MPLS-TP Label Switched Paths (LSPs).

Linear protection in mesh networks - networks with arbitrary interconnectivity between nodes - is described in Section 4.7 of [SurvivFwk]. Linear protection provides rapid and simple protection switching. In a mesh network, linear protection provides a very suitable protection mechanism because it can operate between any pair of points within the network. It can protect against a defect in an intermediate node, a span, a transport path segment, or an end-to-end transport path.

[LinProt] defines a Protection State Coordination (PSC) protocol that supports the different 1+1 and 1:1 architectures described in [SurvivFwk]. The PSC protocol is a single-phased protocol that allows the two endpoints of the protection domain to coordinate the protection switching operation when a switching condition is detected on the transport paths of the protection domain.

This document extends the PSC protocol to allow it to support a protection domain that includes multiple working transport paths that are protected by a single protection transport path. The protection transport path is pre-allocated with resources to transport the traffic normally carried by any one of the working transport paths. This is the architecture described in [SurvivFwk] as 1:n protection, and is the generalization of the 1:1 protection architecture already supported by PSC.

1.1. 1:n Protection architecture

Linear protection switching is a fully allocated survivability mechanism. It is fully allocated in the sense that the route and bandwidth of the protection path is reserved for a set of working paths. For 1:n protection the protection path is allocated to protect any one of n working paths between the two endpoints of the protection domain.

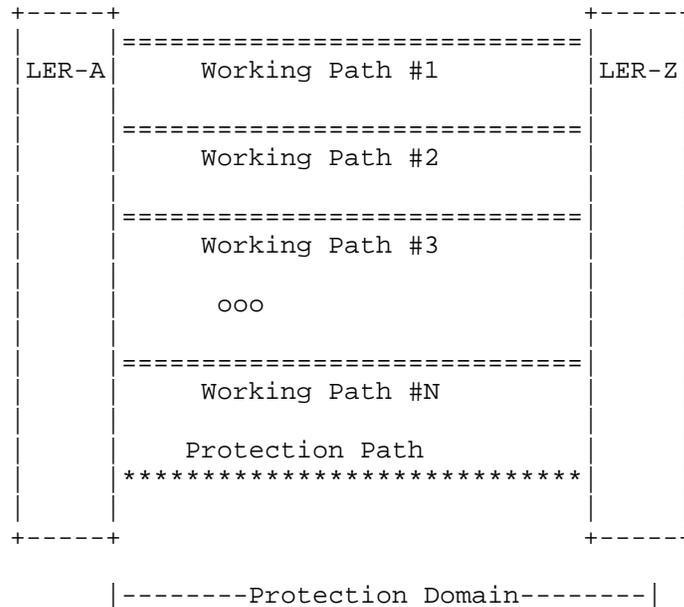


Figure 1: 1:n Protection domain

Figure 1 shows a protection domain with N working transport paths and a single protection path. In 1:n protection, it is assumed (as mentioned above) that the protection path may transport the traffic of only a single working path at any particular time. The identity of the working path that is being protected must be communicated between the two endpoints.

The different working paths may be disjoint at the intermediary points on the path between LER-A and LER-Z and may also have different resource requirements. In addition, each of the working paths may be assigned a priority that could be used to decide which working path would be protected in cases of conflict (see more on this topic in Section 1.3). It is usually advised to arrange these protection groups in a way that would minimize any potential conflict situation.

1.2. Path priority

As the 1:n architecture requires the ability for one working path to preempt the traffic of another in the event of multiple failures (see Section 1.3), there must be an indication of priority between the different working paths so that an implementation can decide whether a new failure should be allowed to preempt a protection switch

already in place. This priority is purely a local decision, i.e., determined by configuration at both endpoints of the protection domain. It is also possible to assign the same priority to multiple working paths, thus creating a "first come first served" preemption policy. This document provides no means to signal the priority of a given working path, nor a means to detect priority mismatches or misconfigurations. Any mismatch or misconfiguration will likely result in unexpected protection behavior.

1.3. Preemption

Preemption occurs when the protection path is being used to transport traffic and is then required to transport traffic for a service with higher priority. At this point, the current traffic that is being transported on the protection path needs to be interrupted to allow the transport of the protected traffic.

There are two basic scenarios for preemption of traffic -

1. When the protection path is used to transport "extra traffic". While this practice is discouraged by [TPReq], it is still not precluded. When the protection domain triggers a protection switch, the extra traffic should be preempted to allow the transport of the protected traffic from the working path that triggered the switching operation. The subsequent treatment of the interrupted service is out of the scope of this document.
2. When the protection path is transporting traffic from a working path and a second working path triggers a switching condition. This second trigger may either be a trigger with a higher priority (e.g. FS after a SF) or because the operator had assigned a higher priority to the working path of the second trigger. At this point, the traffic for the lower priority working path will be interrupted, and the higher priority traffic will be transmitted on the protection path. The preempted traffic will only renew transmission, when either the working path recovers, or the higher priority traffic relinquishes control of the protection path.

1.4. Contributing authors

Nurit Sprecher (NSN)

2. Conventions used in this document

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this

document are to be interpreted as described in [RFC2119].

2.1. Acronyms

This draft uses the following acronyms:

Ack	Acknowledge
DNR	Do not revert
FS	Forced Switch
LER	Label Edge Router
LO	Lockout of protection
MPLS-TP	Transport Profile for MPLS
MS	Manual Switch
NR	No Request
P2P	Point-to-point
P2MP	Point-to-multipoint
PSC	Protection State Coordination Protocol
SD	Signal Degrade
SF	Signal Fail
Wfa	Wait for Acknowledge
WTR	Wait-to-Restore

2.2. Definitions and Terminology

The terminology used in this document is based on the terminology defined in [RFC4427] and further adapted for MPLS-TP in [SurvivFwk]. In addition, we use the term LER to refer to a MPLS-TP Network Element, whether it is a LSR, LER, T-PE, or S-PE.

3. Changes to PSC

The Protection State Coordination protocol (PSC) is defined in [LinProt]. This includes both the format of the G-ACh based message as well as a description of the operations and the state transition logic of the protocol. The extension to cover 1:n protection includes changes to both aspects of PSC.

The changes to the message structure, include both the addition of new information and extension of the semantics of some of the existing fields of the message. These changes will be described in Section 3.2.

The changes relative to the behavior of the base PSC protocol will be described in Section 3.3.

3.1. PSC

Base PSC (as defined in [LinProt]) is a single-phased protocol, i.e. the endpoints perform protection switching without waiting for acknowledgement from the far end LER. The protocol messages are transmitted using the G-ACh and the format is described in Figure 2.

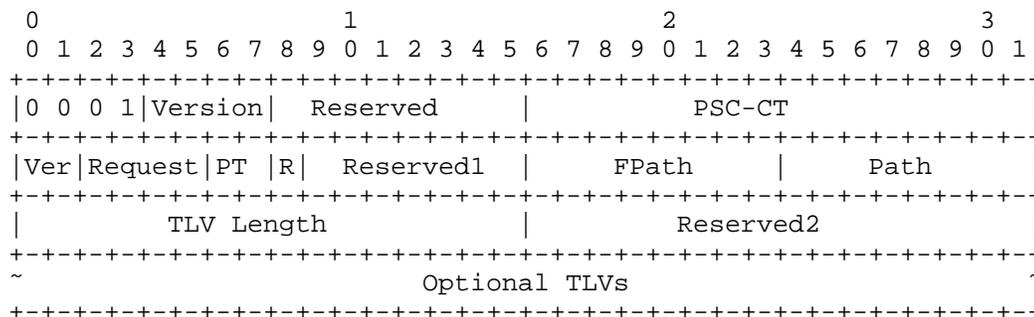


Figure 2: Format of basic PSC packet with a G-ACh header

In regards to the G-ACh Header no changes are suggested in the extensions for 1:n protection, i.e., the channel type field will continue to use the PSC-CT value defined in [LinProt]. The fields from the PSC payload which are affected by this document are the Ver field, the Reserved1 field, and the Fpath and Path fields.

3.2. Changes to PSC Payload

In order to support 1:n protection there is a need to make changes to the format of the PSC payload (see Figure 3). In particular, there is the need to add a new field to the payload to indicate an acknowledge of a protection switching operation. In addition, the semantics of the FPath and Path field are adjusted to indicate an index of the multiple working paths. The details of these changes are supplied in the following subsections.

Due to the significance of these changes, the value of the Ver field (in the PSC payload) for 1:n protection domain MUST be set to 2.

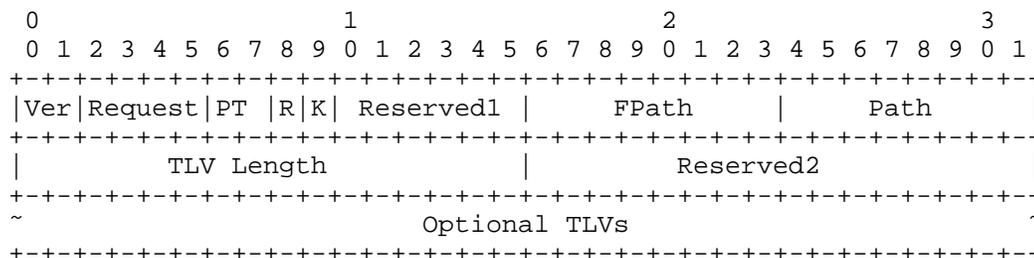


Figure 3: Format of 1:n PSC message payload

3.2.1. Acknowledge (K) flag

The Acknowledge flag is used by an endpoint to acknowledge the request to preempt any current traffic on the protection path and instead transmit the traffic from the requested working path. See details in section x.y.

3.2.2. Fault path (FPath) field

The Fpath field indicates which path is identified to be in a fault condition or affected by an administrative command. The following are the possible values:

- o 0: indicates that the anomaly condition is on the protection path
- o 1-128: indicates that the anomaly condition is on a working path whose index is indicated.
- o 129-255: for future extensions or experimental use.

3.2.3. Data path (Path) field

The Path field indicates which data is being transmitted on the protection path. Under normal conditions, the protection path does not need to carry any user data traffic, but may carry extra traffic. If there is a failure/degrade condition on one of the working paths, then that working path's data traffic will be transmitted over the protection path. The following are the possible values:

- o 0: indicates that the protection path is not transporting user data traffic.
- o 1-128: indicates that the protection path is transmitting user traffic replacing the use of the working path indexed.

- o 129-255: for future extensions or experimental use.

3.3. Changes to PSC Operation

In all of the following subsections, assume a protection domain between LER-A and LER-Z, using working paths 1-N and the protection path as shown in figure 1.

A basic premise of this protection architecture is that both endpoints of the protection domain are configured to associate the indices of the working paths with the proper LSP identifiers. If this condition is not met then the protection scheme will cause inconsistencies in traffic transmission.

3.3.1. Basic operation

Protection of the N working paths is based on the operational principles outlined in [LinProt] and will employ the same basic Protection State Coordination Protocol (PSC) outlined in that document. However, as can be expected, due to certain basic differences in the architecture of the protection domain, a small set of differences in operation are necessary. The following subsections will highlight these differences and explain their effects on the PSC state machine.

3.3.2. Two-phased operation

PSC, as presented in [LinProt] is a single-phased protocol. This means that when an endpoint receives a trigger to perform a protection switch, the LER switches traffic and then notifies the far end of the switch, without waiting for acknowledgement. When addressing the situation in a 1:n protection domain, the endpoint that receives the trigger must first verify that the protection path is available to transmit the protected traffic. This may involve interrupting the traffic that is currently being transmitted on the protection path by both endpoints.

In general, after the LER has detected a trigger for protection switching, e.g. a FS operator command, or a SF indication for one of the working paths, the LER SHALL transmit the appropriate PSC message as described in [LinProt] with the following changes:

- o If the protection domain is currently in either Protecting administrative or Protecting failure state, then the endpoint SHALL verify that the new trigger has a higher priority than the currently protected traffic. If the new trigger has a lower priority then it SHOULD be ignored.

- o The PSC message SHALL set the FPath value to the index of the working path that generated the trigger. The Path value SHOULD be set to 0, unless the protection path was previously transporting traffic from another working path (as indicated by the value of the Path field).
- o If the protection path is currently transporting protected traffic, then the endpoint SHALL block all traffic of the protected working path.
- o The endpoint SHALL transfer to Wfa state (see below).
- o Upon reception of the switching PSC message, the far end LER SHALL verify that the received request is of higher priority than the known current traffic on the protection path, and if so SHALL interrupt the current traffic on the protection path, perform the switch to the requested protected traffic, and send an Acknowledge message (i.e. a PSC message with the Acknowledge flag set to 1) with the Path field set to the index of the protected working path.
- o Upon reception of the Acknowledge message, the initiating LER SHALL perform the protection switch and transmit the appropriate PSC message, with the FPath field indicating the index of the working path that triggered the protection switch and the Path field set to the index of the working path whose traffic is being transported on the protection path.

3.3.3. Acknowledge message

As stated above, before performing a protection switch the endpoint that detected a switching trigger MUST wait for an Acknowledge message prior to performing the switch. There are two types of message that will be considered as an Acknowledge message:

1. A reply message with the Request field reflecting the state of the far end, and the Path field set to the index of the working path that triggered the switching condition. For example, if there is a Forced Switch command detected by LER-Z on working path #4, then LER-Z will have sent an FS(4,0) message to LER-A. Then when LER-Z receives a message such as NR(0,4)Ack this should be considered acknowledgement of the switching and that the protection path is available to switch the traffic from working path #4.
2. A remote message with the same Request field and FPath field as that transmitted by the LER in the Wfa state. For example, if there is a bi-directional Signal fault detected by LER-A on

working path #2, then LER-A will enter Wfa state and transmit a SF(4,0) message. When it receives the SF(4,0) message from LER-Z, that has also detected the SF condition, it should be considered an acknowledgement of the switching and that the protection path is available to switch the traffic from working path #2.

3.3.4. Wait for Acknowledge (Wfa) timer

The protection system should include a timer called the Wait for Acknowledge (Wfa) timer that SHALL be started when the LER enters Wfa state and reset when the Acknowledge message is received. The length of the Wfa timer SHOULD be configured to allow protection switching within the normal time constraints. The Wfa timer will expire only if no Acknowledge message was received by the LER in Wfa state. The Wfa Expires local input should have a priority just below that of the WTRExpires signal.

3.3.5. Additional PSC State

As described above, there is a need for the endpoint that is reporting on a trigger for protection-switching to delay the actual switchover until an acknowledge is received from the far end LER. In order to facilitate this wait period it is necessary to define a new PSC State - Wait for Acknowledge (Wfa) state. This state will be entered by the LER upon receiving a trigger for protection switching, and will be exited either upon receiving an acknowledge message or receiving a remote message indicating that the protection path is currently occupied by a higher priority request.

The following sub-section will describe the actions to be taken when an LER is in the Wfa state.

3.3.5.1. Wait for Acknowledge (Wfa) State

An LER will enter the Wait for Acknowledge state before transitioning into a protection state, i.e. either Protecting administrative or Protecting failure state. The LER SHALL remain in this state until either receiving an Acknowledge message, or until a Wfa timer expires. Normally, the Acknowledge message will be a remote PSC input. The following describe how the LER, in Wfa state, should react to a new local input:

- o A local Clear SHALL cause the LER to go into Normal state if the LER is in Wfa state due to either a FS or MS trigger and transmit an NR(0,0) PSC message. If the LER is in Wfa state due to a SF trigger then the local Clear SHALL be ignored.

- o A local LO SHALL cause the LER to go into Unavailable state and begin to transmit LO(x, 0) [where x indicates the index of the working path that triggered the Wfa state].
- o A local FS SHALL cause the LER to remain in Wfa state and transmit the FS(x, 0) message [where x indicates the index of the protected working path]. If the LER is in Wfa state due to a FS from a different working path, then the working path with the higher priority SHALL be the protected working path. If the LER is in Wfa state due to any other switching trigger, then the working path that is identified in this FS will be the protected working path.
- o A local SF SHALL cause the LER to remain in Wfa state. If the LER is in Wfa state due to an existing FS trigger, then ignore the local SF and continue to transmit the FS(x, 0) PSC message. If the LER is in Wfa state due to an existing SF trigger then transmit the SF(x, 0) PSC message [where x indicates the index of protected working path, i.e. the highest priority working path indicating an SF condition]. If the LER is in Wfa state due to any other trigger, then begin transmitting a SF(x, 0) PSC message [where x indicates the index of the working path that is generating the SF condition].
- o A local ClearSF indication where the working path is the same as the path that triggered the LER into Wfa state SHALL cause the LER to go into WTR state (note: 1:N protection is always revertive) and to transmit the WTR(0, 0) message. If the ClearSF indicates a different index from the protected working path or incates the protection path then the indication SHALL be ignored.
- o A local MS operator command SHALL cause the LER to remain in Wfa state. If the LER is in Wfa state due an existing MS trigger, then the node continues to transmit MS(x, 0) messages [where x indicates the index of the protected working path, i.e. the highest priority working path indicating the MS condition]. If the LER is in Wfa state due to any other trigger, ignore the MS command and continue transmitting the current message.
- o If the Wfa timer expires, i.e. the LER did not receive the Acknowledge message from the far end in a timely manner, then the LER SHALL go to Unavailable state, i.e. it assumes that there is a problem on the protection path (where all PSC traffic is transmitted) and send an error notification to the management system. The LER SHALL continue transmitting the current PSC message with Path field set to 0.

- o All other local indications SHALL be ignored.

The following details the reactions of the LER in Wfa state to remote messages:

- o Any remote message with the Acknowledge flag set to 1 and the Path field set to the index of the protected working path SHALL cause the LER to change state. If the trigger was either FS or MS command, the LER enters Protecting administrative state. The LER transmits the appropriate message according to the trigger (i.e. FS(x,x) for FS command and MS(x,x) for the MS command). If the trigger was a SF condition, then the LER enters the Protecting failure state and begins to transmit the appropriate SF(x, x) message. A remote message with the Acknowledge flag set to 1 but where the Path field does not match, according to the description above, SHALL be ignored.
- o A remote LO message SHALL cause the LER to go into Unavailable state and transmit the appropriate message for the trigger that caused the Wfa state.
- o A remote FS message indicating the same working path as the local FS command that triggered the Wfa state SHALL be considered an Acknowledge message, even if the Acknowledge flag is not set. The LER SHALL perform the protection switch, and begin transmitting the FS(x, x) message [where x indicates the index of the protected working path]. If the remote FS message indicates a different index than the one indicated in the local FS and if the remote FS message indicates a lower priority working path than the working path in the local FS trigger then the LER SHALL ignore the remote FS message and remain in Wfa state. If the remote FS message indicates an index of higher priority or the LER is in Wfa state as a result of a SF or MS trigger, then the LER SHALL perform the protection switch for the protected working path indicated by the remote FS message, and SHALL go to Protecting administrative state and transmit the appropriate message for the local trigger with the Path field set to the index of the remote message and the Acknowledge flag set to 1.
- o A remote SF message indicating an error on the protection path SHALL cause the LER to go into Unavailable state and transmit the appropriate message for the trigger that caused to Wfa state.
- o A remote SF message indicating an error on the same working path as the local SF condition that triggered the Wfa state SHALL be considered an Acknowledge message (even if the Acknowledge flag is not set). The LER SHALL perform the protection switch, go to Protecting failure state and transmit the SF(x, x) message [where

x is the index of the protected working path]. If the remote SF message indicates a different index than the one indicated in the local SF, then if the local command indicates a higher priority working path the LER SHALL ignore the remote SF message and remain in Wfa state. If the remote SF message indicates an index of higher priority or the LER is in Wfa state as a result of a MS trigger, then the LER SHALL perform the protection switch for the protected working path indicated by the remote SF message, and SHALL go to Protecting failure state and transmit the appropriate message for the local trigger with the Path field set to the index of the remote message and the Acknowledge flag set to 1. If the LER is in Wfa state due to a local FS command, then it SHALL ignore the remote message and remain in Wfa state.

- o A remote MS message indicating an error on the same working path as the local MS that triggered the Wfa state SHALL be considered an Acknowledge message (even if the Acknowledge flag is not set). The LER SHALL perform the protection switch, go to Protecting administrative state and transmit the MS(x, x) message [where x is the index of the protected working path]. If the remote MS message indicates a different index than the one indicated in the local MS, then if the local command indicates a higher priority working path or the LER is in Wfa due to either a FS or SF trigger, the LER SHALL ignore the remote MS message and remain in Wfa state. If the remote MS message indicates an index of higher priority, then the LER SHALL perform the protection switch for the protected working path indicated by the remote MS message, and SHALL go to Protecting administrative state and transmit an NR(0, y) with the Path field set to the index of the remote message and the Acknowledge flag set to 1.

- o All other remote messages SHOULD be ignored.

4. IANA Considerations

This document does not include any required IANA considerations

5. Security Considerations

The generic security considerations for the data-plane of MPLS-TP are described in the security framework document [SecureFwk] together with the required mechanisms needed to address them. The security considerations for the generic associated control channel are described in [RFC5586]. The security considerations for protection and recovery aspects of MPLS-TP are addressed in [SurvivFwk].

The extensions to the protocol described in this document are extensions to the protocol defined in [LinProt] and does not introduce any new security risks.

6. Acknowledgements

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

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.
- [TPReq] Niven-Jenkins, B., Brungard, D., Betts, M., Sprecher, N., and S. Ueno, "Requirements of an MPLS Transport Profile", RFC 5654, September 2009.
- [LinProt] Bryant, S., Sprecher, N., Osborne, E., Fulignoli, A., and Y. Weingarten, "Multi-protocol Label Switching Transport Profile Linear Protection", ID draft-ietf-mpls-tp-linear-protection-07.txt, Apr 2011.

7.2. Informative References

- [RFC5586] Vigoureux,, M., Bocci, M., Swallow, G., Aggarwal, R., and D. Ward, "MPLS Generic Associated Channel", RFC 5586, May 2009.
- [RFC4427] Mannie, E. and D. Papadimitriou, "Recovery Terminology for Generalized Multi-Protocol Label Switching", RFC 4427, Mar 2006.
- [SurvivFwk] Sprecher, N., Farrel, A., and H. Shah, "Multi-protocol Label Switching Transport Profile Survivability Framework", ID draft-ietf-mpls-tp-survive-fwk-02.txt, Feb 2009.
- [SecureFwk] Fang, L., Niven-Jenkins, B., Mansfield, S., Zhang, R., Bitar, N., Daikoku, M., and L. Wang, "MPLS-TP Security

Framework",
 ID draft-ietf-mpls-tp-security-framework-00.txt, Feb 2011.

Appendix A. PSC state machine tables

The full PSC state machine is described in [LinProt], both in textual and tabular form. This appendix highlights the changes to the basic PSC state machine. In the event of a mismatch between these tables and the text either in [LinProt] or in this document, the text is authoritative. Note that this appendix is intended to be a functional description, not an implementation specification.

The tables here use the same format and state descriptions used in the Linear Protection document with the addition of the Wfa state, Wfa Expires, and the changes in the behavior that is noted.

Each state corresponds to the transmission of a particular set of Request, FPath and Path bits. The table below lists the message that is generally sent in each particular state. If the message to be sent in a particular state deviates from the table below, it is noted in the footnotes to the state-machine table.

State	REQ(FP,P)
-----	-----
N	NR(0,0)
UA:LO:L	LO(0,0)
UA:P:L	SF(0,0)
UA:LO:R	NR(0,0)
UA:P:R	NR(0,0)
PF:W:L	SF(1,1)
PF:W:R	NR(0,1)
PA:F:L	FS(1,1)
PA:M:L	MS(1,1)
PA:F:R	NR(0,1)
PA:M:R	NR(0,1)
WTR	WTR(0,1)
DNR	DNR(0,1)

The top row in each table is the list of possible inputs. The local inputs are:

NR No Request
OC Operator Clear
LO Lockout of protection
SF-P Signal Fail on protection path
SF-W Signal Fail on working path
FS Forced Switch
SFc Clear Signal Fail
MS Manual Switch
WTRExp WTR Expired

and the remote inputs are:

LO remote LO message
SF-P remote SF message indicating protection path
SF-W remote SF message indicating working path
FS remote FS message
MS remote MS message
WTR remote WTR message
DNR remote DNR message
NR remote NR message

Section 4.3.3 refers to some states as 'remote' and some as 'local'. By definition, all states listed in the table of local sources are local states, and all states listed in the table of remote sources are remote states. For example, section 4.3.3.1 says "A local Lockout of protection input SHALL cause the LER to go into local Unavailable State". As the trigger for this state change is a local one, 'local Unavailable State' is by definition displayed in the table of local sources. Similarly, "A remote Lockout of protection message SHALL cause the LER to go into remote Unavailable state" means that the state represented in the Unavailable rows in the table of remote sources is by definition a remote Unavailable state.

Each cell in the table below contains either a state, a footnote, or the letter 'i'. 'i' stands for Ignore, and is an indication to continue with the current behavior. See section 4.3.3. The footnotes are listed below the table.

Part 1: Local input state machine

	OC	LO	SF-P	FS	SF-W	SF _c	MS	WTRExp
N	i	UA:LO:L	UA:P:L	PA:F:L	PF:W:L	i	PA:M:L	i
UA:LO:L	N	i	i	i	i	i	i	i
UA:P:L	i	UA:LO:L	i	i	i	[5]	i	i
UA:LO:R	i	UA:LO:L	[1]	i	[2]	[6]	i	i
UA:P:R	i	UA:LO:L	UA:P:L	i	[3]	[6]	i	i
PF:W:L	i	UA:LO:L	UA:P:L	PA:F:L	i	[7]	i	i
PF:W:R	i	UA:LO:L	UA:P:L	PA:F:L	PF:W:L	i	i	i
PA:F:L	N	UA:LO:L	UA:P:L	i	i	i	i	i
PA:M:L	N	UA:LO:L	UA:P:L	PA:F:L	PF:W:L	i	i	i
PA:F:R	i	UA:LO:L	UA:P:L	PA:F:L	[4]	[8]	i	i
PA:M:R	i	UA:LO:L	UA:P:L	PA:F:L	PF:W:L	i	PA:M:L	i
WTR	i	UA:LO:L	UA:P:L	PA:F:L	PF:W:L	i	PA:M:L	[9]
DNR	i	UA:LO:L	UA:P:L	PA:F:L	PF:W:L	i	PA:M:L	i

Part 2: Remote messages state machine

	LO	SF-P	FS	SF-W	MS	WTR	DNR	NR
N	UA:LO:R	UA:P:R	PA:F:R	PF:W:R	PA:M:R	i	i	i
UA:LO:L	i	i	i	i	i	i	i	i
UA:P:L	[10]	i	i	i	i	i	i	i
UA:LO:R	i	i	i	i	i	i	i	[16]
UA:P:R	UA:LO:R	i	i	i	i	i	i	[16]
PF:W:L	[11]	[12]	PA:F:R	i	i	i	i	i
PF:W:R	UA:LO:R	UA:P:R	PA:F:R	i	i	[14]	[15]	N
PA:F:L	UA:LO:R	UA:P:R	i	i	i	i	i	i
PA:M:L	UA:LO:R	UA:P:R	PA:F:R	[13]	i	i	i	i
PA:F:R	UA:LO:R	UA:P:R	i	i	i	i	i	[17]
PA:M:R	UA:LO:R	UA:P:R	PA:F:R	[13]	i	i	i	N
WTR	UA:LO:R	UA:P:R	PA:F:R	PF:W:R	PA:M:R	i	i	[18]
DNR	UA:LO:R	UA:P:R	PA:F:R	PF:W:R	PA:M:R	i	i	i

The following are the footnotes for the table:

[1] Remain in the current state (UA:LO:R) and transmit SF(0,0)

[2] Remain in the current state (UA:LO:R) and transmit SF(1,0)

[3] Remain in the current state (UA:P:R) and transmit SF(1,0)

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

[5] If the SF being cleared is SF-P, Transition to N. If it's SF-W, ignore the clear.

- [6] Remain in current state (UA:x:R), if the SFc corresponds to a previous SF then begin transmitting NR(0,0).
- [7] If domain configured for revertive behavior transition to WTR, else transition to DNR
- [8] Remain in PA:F:R and transmit NR(0,1)
- [9] Remain in WTR, send NR(0,1)
- [10] Transition to UA:LO:R continue sending SF(0,0)
- [11] Transition to UA:LO:R and send SF(1,0)
- [12] Transition to UA and send SF(1,0)
- [13] Transition to PF:W:R and send NR(0,1)
- [14] Transition to WTR state and continue to send the current message.
- [15] Transition to DNR state and continue to send the current message.
- [16] If the local input is SF-P then transition to UA:P:L. If the local input is SF-W then transition to PF:W:L. Else - transition to N state and continue to send the current message.
- [17] If the local input is SF-W then transition to PF:W:L. Else - transition to N state and continue to send the current message.
- [18] If the receiving LER's WTR timer is running, maintain current state and message. If the WTR timer is stopped, transition to N.

Authors' Addresses

Eric Osborne
Cisco
United States

Email: eosborne@cisco.com

Fei Zhang
ZTE
China

Email: zhang.fei3@zte.com.cn

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

Email: yaacov.weingarten@nsn.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/lid-abstracts.txt>.

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

This Internet-Draft will expire on January 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

1. Introduction.....	4
1.1. Background and Motivation.....	4
1.2. Contributing authors.....	5
2. Terminologies.....	5
3. Overview of MPLS-TP base functions.....	6
3.1. MPLS-TP development principles.....	6
3.2. Data Plane.....	7
3.3. Control Plane.....	7
3.4. OAM.....	7
3.5. Survivability.....	8
4. MPLS-TP Use Case Studies.....	8
4.1. Mobile Backhaul.....	8
4.2. Metro Access and Aggregation.....	10
4.3. Packet Optical Transport.....	10
5. Network Design Considerations.....	11
5.1. IP/MPLS vs. MPLS-TP.....	11
5.2. Standards compliance.....	11
5.3. End-to-end MPLS OAM consistency.....	12
5.4. Delay and delay variation.....	12
5.5. General network design considerations.....	15
6. MPLS-TP Deployment Consideration.....	15
6.1. Network Modes Selection.....	15
6.2. Provisioning Modes Selection.....	16
7. Security Considerations.....	16
8. IANA Considerations.....	16
9. Normative References.....	17
10. Informative References.....	17
11. Author's Addresses.....	17

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
- Resiliency
- 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 Foredbu
Email: Lai.wang@telenor.com

Mach(Guoyi) Chen
Huawei Technologies Co., Ltd.
No. 3 Xixi 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

Network Working Group
Internet-Draft
Intended status: Informational
Expires: January 12, 2012

A. Farrel
Huawei Technologies
H. Endo
Hitachi, Ltd.
R. Winter
NEC
Y. Koike
NTT
M. Paul
Deutsch Telekom
July 11, 2011

Handling MPLS-TP OAM Packets Targeted at Internal MIPs
draft-farrel-mpls-tp-mip-mep-map-04

Abstract

The Framework for Operations, Administration and Maintenance (OAM) within the MPLS Transport Profile (MPLS-TP) describes how Maintenance Entity Group Intermediate Points (MIPs) may be situated within network nodes at the incoming and outgoing interfaces.

This document describes a way of forming OAM messages so that they can be targeted at MIPs on incoming or MIPs on outgoing interfaces, forwarded correctly through the "switch fabric", and handled efficiently in node implementations where there is no distinction between the incoming and outgoing MIP.

The material in this document is provided for discussion within the MPLS-TP community in the expectation that this idea or some similar mechanism will be subsumed into a more general MPLS-TP OAM document.

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

Status of this Memo

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

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

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

This Internet-Draft will expire on January 12, 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	4
2. Requirements notation	5
3. Terminology	6
4. Summary of the Problem Statement	7
5. Overview	10
5.1. Rejected Partial Solution	12
6. Possible Solutions	14
6.1. ID-based Solution	14
6.2. Using an ACH reserved bit	15
7. Security Considerations	17
8. IANA Considerations	18
9. Acknowledgments	19
10. References	20
10.1. Normative References	20
10.2. Informative References	20
Appendix A. Previously considered solutions	21
A.1. GAL TTL	21
A.2. A separate channel type for the out-MIP	21
A.3. Decrement TTL once per MIP	21
Authors' Addresses	22

1. Introduction

The Framework for Operations, Administration and Maintenance (OAM) within the MPLS Transport Profile (MPLS-TP) (the MPLS-TP OAM Framework, [I-D.ietf-mpls-tp-oam-framework]) distinguishes two configurations for Maintenance Entity Group Intermediate Points (MIPs) on a node. It defines per-node MIPs and per-interface MIPs, where a per-node MIP is a single MIP per node in an unspecified location within the node and per-interface MIPs are two (or more) MIPs per node on both sides of the forwarding engine.

In-band OAM messages are sent using the Generic Associated Channel (G-ACh) [RFC5586]. OAM messages for the transit points of pseudowires (PWs) or Label Switched Paths (LSPs) are delivered using the expiration of the MPLS shim header time-to-live (TTL) field. OAM messages for the end points of PWs and LSPs are simply delivered as normal.

OAM messages delivered to end points or transit points are distinguished from other (data) packets so that they can be processed as OAM. In LSPs, the mechanism used is the presence of the Generic Associated Channel Label (GAL) in the Label Stack Entry (LSE) under the top LSE [RFC5586]. In PWs, the mechanism used is the presence of the PW Associated Channel Header (PWACH) [RFC4385].

In case multiple MIPs are present on a single node, these mechanisms alone provide no way to address one particular MIP out of the set of MIPs.

This document describes a way of forming OAM messages so that they can be targeted at incoming MIPs and outgoing MIPs, forwarded correctly through the "switch fabric", and handled efficiently in node implementations where there is no distinction between the incoming and outgoing MIP.

The material in this document is provided for discussion within the MPLS-TP community in the expectation that this idea or some similar mechanisms will be subsumed into a more general MPLS-TP OAM document.

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

Note that the acronym "OAM" is used in conformance with [RFC6291].

2. Requirements notation

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

3. Terminology

In this document we use the term in-MIP (incoming MIP) to refer to the MIP which processes OAM messages before they pass through the forwarding engine of a node. An out-MIP (outgoing MIP) processes OAM messages after they have passed the forwarding engine of the node. The two together are referred to as internal MIPs.

4. Summary of the Problem Statement

Figure 1 shows an abstract functional representation of an MPLS-TP node. It is decomposed as an incoming interface, a cross-connect (XC), and an outgoing interface. As per the discussion in [I-D.ietf-mpls-tp-oam-framework], MIPs may be placed in each of the functional interface components.

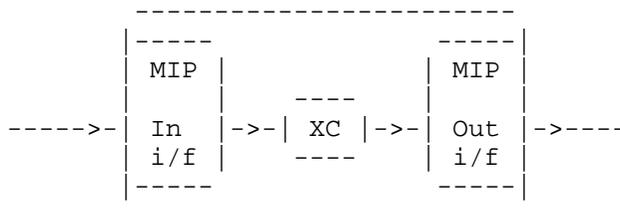


Figure 1: Abstract Functional Representation of an MPLS-TP Node

Several distinct OAM functions are required within this architectural model such as:

- o CV between a MEP and a MIP
- o traceroute over an MPLS-TP LSP and/or an MPLS-TP PW containing MIPs
- o OAM control at a MIP
- o data-plane loopback at a MIP
- o diagnostic tests

The MIPs in these OAM functions may equally be the MIPs at the incoming or outgoing interfaces.

Per-interface MIPs have the advantage that they enable a more accurate localization and identification of faults and targeted performance monitoring or diagnostic test. In particular, the identification of whether a problem is located between nodes or on a particular node and where on that node is greatly enhanced. For obvious reasons, it is important to narrow the cause of a fault down quickly to initiate a timely, and well-directed maintenance action to resume normal network operation.

The following two figures illustrate the fundamental difference of using per-node and per-interface MEPs and MIPs for OAM. Figure 2 depicts OAM using per-interface MIPs and MEPs. For reasons of

which gives operators better information to deal with adverse networking conditions.

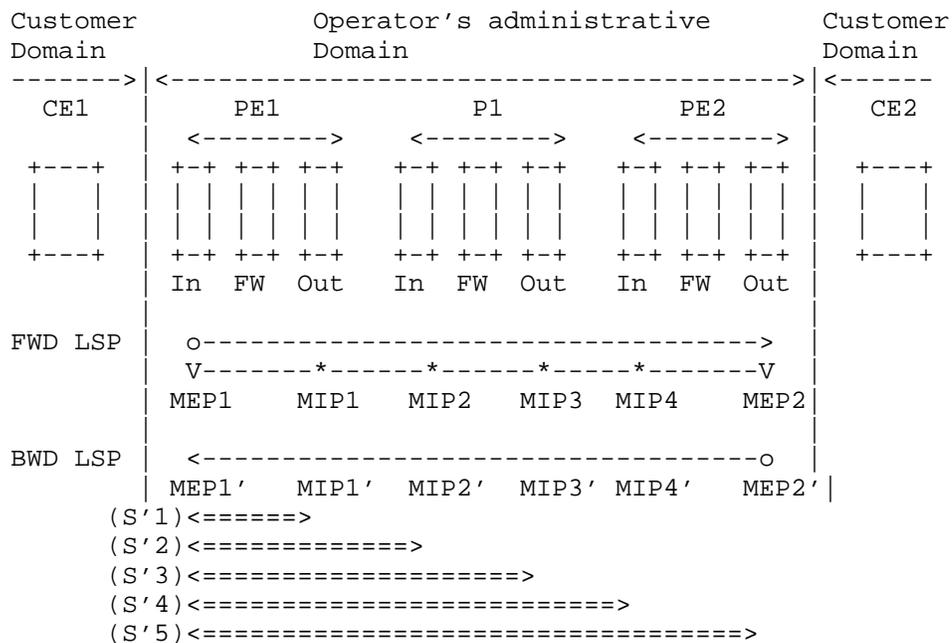


Figure 3: Example of OAM relying on per-interface MIPs and MEPs

5. Overview

In-band OAM messages are sent using the G-ACh [RFC5586] for MPLS-TP LSPs and MPLS-TP PWs, respectively. OAM messages for the transit points of PWs or LSPs are delivered using the expiration of the time-to-live (TTL) field in the top LSE of the MPLS packet header. OAM messages for the end points of PWs and LSPs are simply delivered as normal.

OAM messages delivered to end points or transit points are distinguished from other (data) packets so that they can be processed as OAM. In LSPs, the mechanism used is the presence of the Generic Associated Channel Label (GAL) in the LSE under the top LSE [RFC5586]. In PWs, the mechanism used is the presence of the PW Associated Channel Header [RFC4385].

Any solution for sending OAM to the in and out-MIPs must fit within these existing models of handling OAM.

Additionally, many MPLS-TP nodes contain an optimization such that all queuing and the forwarding function is performed at the incoming interface. The abstract functional representation of such a node is shown in Figure 4. As shown in the figure, the outgoing interfaces are minimal and for this reason it may not be possible to include MIP functions on those interfaces. This is in particular the case for existing deployed implementations.

Any solution that attempts to send OAM to the outgoing interface of an MPLS-TP node must not cause any problems when such implementations are present.

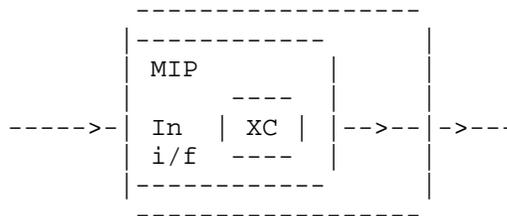


Figure 4: Abstract Functional Representation of an Optimized MPLS-TP Node

Lastly, OAM must operate on MPLS-TP nodes that are branch points on point-to-multipoint (P2MP) trees. That means that it must be possible to target individual outgoing MIPs as well as all outgoing MIPs in the abstract functional representation shown in Figure 5, as

well as to handle the optimized P2MP node as shown in Figure 6.

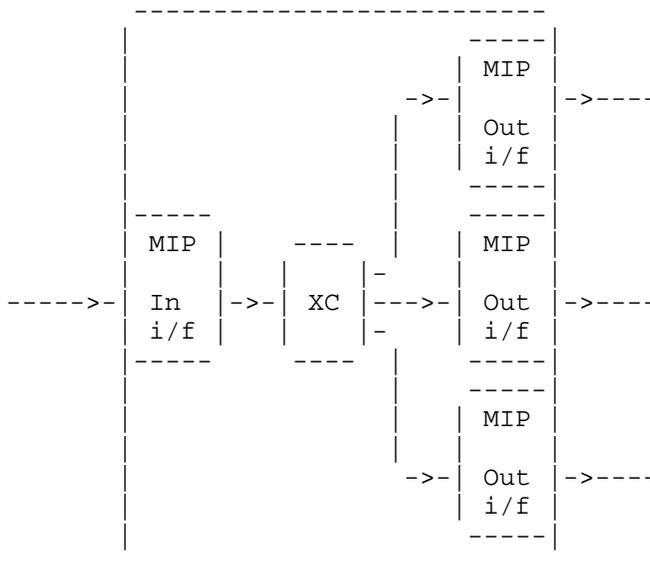


Figure 5: Abstract Functional Representation of an MPLS-TP Node Supporting P2MP

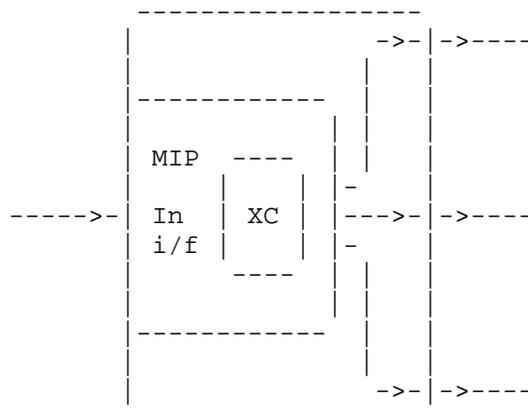


Figure 6: Abstract Functional Representation of an Optimized MPLS-TP Node Supporting P2MP

In summary, the solution for OAM message delivery must support the

following features:

- o Forwarding of OAM packets exactly as data packets.
- o Delivery of OAM messages to the correct MPLS-TP node.
- o Direction of OAM instructions to the correct MIP within an MPLS-TP node.
- o Packet inspection at the incoming and outgoing interfaces must be minimized.

Note that although this issue appears superficially to be an implementation matter local to an individual node, the format of the message needs to be standardised so that:

- o An upstream MEP can correctly target the outgoing MIP of a specific MPLS-TP node.
- o A downstream node can correctly filter out any OAM messages that were intended for its upstream neighbor's outgoing MIP, but which were not handled there because the upstream neighbor is an optimized implementation.

Note that the last bullet point describes a safety net and an implementation should avoid that this situation ever arises.

5.1. Rejected Partial Solution

A reject solution is depicted in Figure 7. All data and non-local OAM is handled as normal. Local OAM is intercepted at the incoming interface and delivered to the MIP at the incoming interface. If the OAM is intended for the incoming MIP it is handled there with no issue. If the OAM is intended for the outgoing MIP it is forwarded to that MIP using some internal messaging system that is implementation-specific.

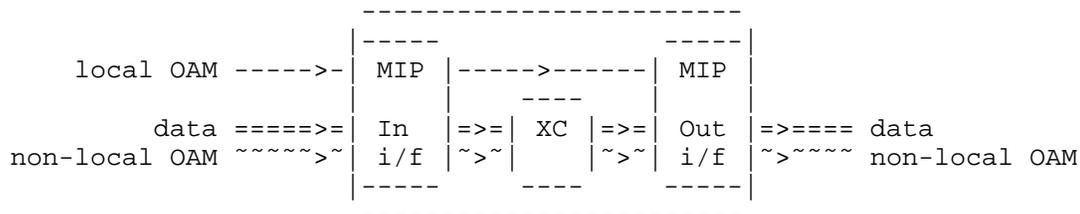


Figure 7: OAM Control Message Delivery Bypassing the Switching Fabric

This solution is fully functional for the incoming MIP. It also supports control of data loopback for the outgoing MIP, and can adequately perform some OAM features such as interface identity reporting at the outgoing MIP.

However, because the OAM message is not forwarded through the switch fabric, this solution cannot correctly perform OAM loopback, connectivity verification, LSP tracing, or performance measurement.

6. Possible Solutions

We briefly present here a number of possible solutions to the problem outlined so far with the hope that the WG will quickly converge towards adopting one of them. The appendix of this document already contains a few solutions that the authors have discarded which have been left in the document for informational purposes.

6.1. ID-based Solution

An ID-based solution leverages existing identification information in OAM messages. OAM solutions therefore need to individually make sure that enough of that information is present to support the per-interface model. In particular, the MIP identifiers as described in [I-D.ietf-mppls-tp-identifiers] need to be present in OAM messages. [I-D.ietf-mppls-tp-identifiers] defines a format that supports the per-interface model which is sufficient for this purpose. In addition, some constraints must be agreed on.

From a requirements-perspective this means:

- o Forwarding of OAM packets exactly as data packets - This way of internal-MIP addressing has no implications on the way data packets and non-local OAM packets are handled. The TTL processing remains untouched. This also means that the TTL will expire on the ingress.
- o Delivery of OAM messages to the correct MPLS-TP node - The TTL addresses the node.
- o Direction of OAM instructions to the correct MIP within an MPLS-TP node - The ID information contained in the OAM packet is used to tell whether the packet is for the in or out-MIP.
- o Packet inspection at the incoming and outgoing interfaces must be minimized - packet inspection becomes a bit more complicated since the required information can be in different places for different types of OAM.
- o An upstream MEP can correctly target the outgoing MIP of a specific MPLS-TP node - this is simple as the TTL addresses the node and the ID information in the packet addresses the respective MIP.
- o A downstream node can correctly filter out any OAM messages that were intended for its upstream neighbor's outgoing MIP, but which were not handled there because the upstream neighbor is an optimized (legacy) implementation - OAM messages expire on the

ingress so the legacy upstream neighbor will process the packet. Since the ID information is not correct, the node will discard the packet. Leakage should therefore not occur.

6.2. Using an ACH reserved bit

The ACH contains eight reserved bits which currently all need to be set to zero and ignored on reception. One bit could be reserved as an out-MIP address flag. In other words, in case the bit is set, the out-MIP is addressed. An advantage of this approach is that there is no semantic overlap with anything that exists today, as the bits are not in use. Existing implementations need to ignore it. That means that existing implementations will process the OAM packets at the in-MIP/per-node MIP.

From a requirements-perspective this means:

- o Forwarding of OAM packets exactly as data packets - This way of internal-MIP addressing has no implications on the way data packets and non-local OAM packets are handled. The TTL processing remains untouched.
- o Delivery of OAM messages to the correct MPLS-TP node - The TTL addresses the node.
- o Direction of OAM instructions to the correct MIP within an MPLS-TP node - The newly defined bit addresses the correct place within the node (0 = in-MIP and 1 = out-MIP).
- o Packet inspection at the incoming and outgoing interfaces must be minimized - packet inspection requires to check an additional bit, which however is at a fixed location.
- o An upstream MEP can correctly target the outgoing MIP of a specific MPLS-TP node - this is simple as the TTL addresses the node and the new flag indicates the place within the node.
- o A downstream node can correctly filter out any OAM messages that were intended for its upstream neighbor's outgoing MIP, but which were not handled there because the upstream neighbor is an optimized (legacy) implementation - Since the TTL will expire on the node the message will be processed by it. Since it is not targeted at that MIP, it will discard it.

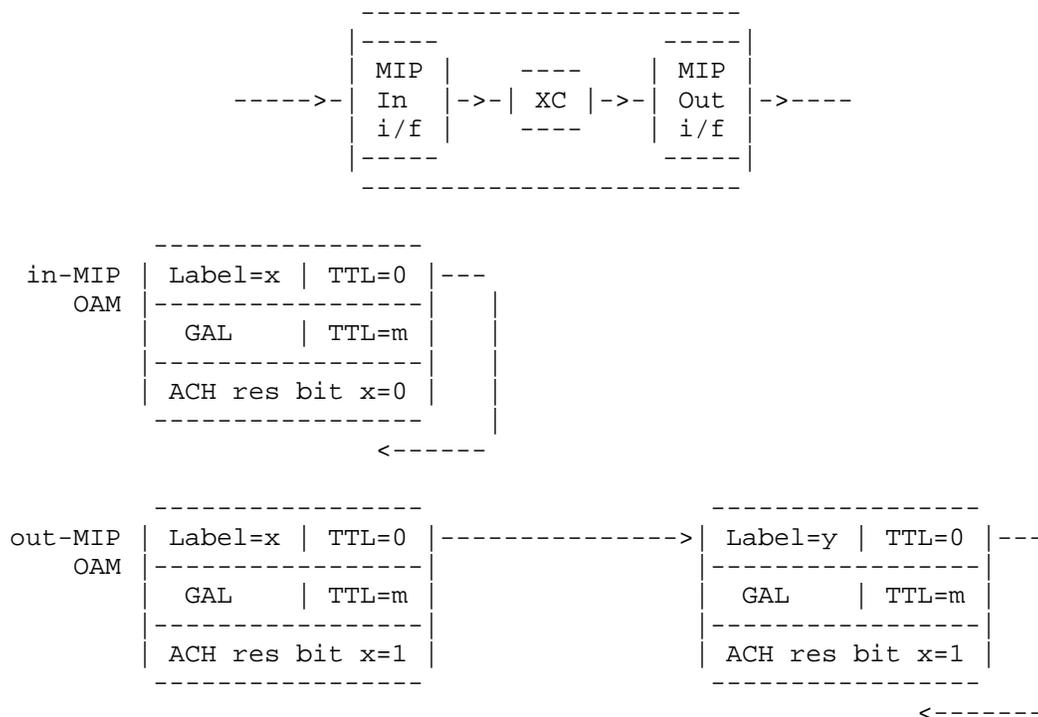


Figure 8: Packet Formats for in and out-MIP OAM (for LSPs)

7. Security Considerations

OAM security is discussed in [I-D.ietf-mpls-tp-oam-framework] and [I-D.manral-mpls-tp-oam-security-tlv].

OAM can provide useful information for detecting and tracing security attacks.

OAM can also be used to illicitly gather information or for denial of service attacks and other types of attack. Implementations therefore are required to offer security mechanisms for OAM. Deployments are strongly advised to use such mechanisms.

Mixing of per-node and per-interface OAM on a single node is not advised as OAM message leakage could be the result.

8. IANA Considerations

This revision of this document does not make any requests of IANA.

9. Acknowledgments

The authors gratefully acknowledge the substantial contributions of Zhenlong Cui. We would also like to thank Eric Gray and Sami Boutros for interesting input to this document through discussions.

10. References

10.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC4385] Bryant, S., Swallow, G., Martini, L., and D. McPherson, "Pseudowire Emulation Edge-to-Edge (PWE3) Control Word for Use over an MPLS PSN", RFC 4385, February 2006.
- [RFC5586] Bocci, M., Vigoureux, M., and S. Bryant, "MPLS Generic Associated Channel", RFC 5586, June 2009.

10.2. Informative References

- [I-D.ietf-mpls-tp-identifiers]
Bocci, M., Swallow, G., and E. Gray, "MPLS-TP Identifiers", draft-ietf-mpls-tp-identifiers-06 (work in progress), June 2011.
- [I-D.ietf-mpls-tp-oam-framework]
Allan, D., Busi, I., Niven-Jenkins, B., Fulignoli, A., Hernandez-Valencia, E., Levrau, L., Sestito, V., Sprecher, N., Helvoort, H., Vigoureux, M., Weingarten, Y., and R. Winter, "Operations, Administration and Maintenance Framework for MPLS-based Transport Networks", draft-ietf-mpls-tp-oam-framework-11 (work in progress), February 2011.
- [I-D.manral-mpls-tp-oam-security-tlv]
Manral, V., "MPLS-TP General Authentication TLV for G-ACH", draft-manral-mpls-tp-oam-security-tlv-00 (work in progress), June 2009.
- [RFC6291] Andersson, L., van Helvoort, H., Bonica, R., Romascanu, D., and S. Mansfield, "Guidelines for the Use of the "OAM" Acronym in the IETF", BCP 161, RFC 6291, June 2011.

Appendix A. Previously considered solutions

A.1. GAL TTL

The use of the GAL TTL has been considered before. This transforms the GAL TTL into some kind of node-internal TTL, i.e. a GAL TTL of 0 would address the in-MIP and a GAL TTL of 1 the out-MIP. The main drawback of this approach is that it (as of now at least) would only be applicable to LSPs and not to PWs.

A.2. A separate channel type for the out-MIP

This approach would require two channel types for the exact same OAM type, one to address the in-MIP and another one to address the out-MIP. This seems like a waste of channel types, however it appears that there is no expected shortage of them. Legacy nodes will discard the packets as the new channel types are unknown. Having two channel types for the same OAM however feels a bit hacky.

A.3. Decrement TTL once per MIP

Decrementing the TTL more than once per node seems a "natural" way of per-interface MIP addressing since TTL expiry is all that is needed for the per-node MIP case. In other words, by decrementing the TTL once per MIP (twice per node) no extra mechanism is needed to solve the internal MIP addressing problem. The solution has been discarded since it does not represent the typical mode of network operation today (since also for normal data packets the TTL needs to be decremented more than once).

Authors' Addresses

Adrian Farrel
Huawei Technologies

Email: adrian.farrel@huawei.com

Hideki Endo
Hitachi, Ltd.

Email: hideki.endo.es@hitachi.com

Rolf Winter
NEC

Email: rolf.winter@neclab.eu

Yoshinori Koike
NTT

Email: koike.yoshinori@lab.ntt.co.jp

Manuel Paul
Deutsch Telekom

Email: Manuel.Paul@telekom.de

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: January 5, 2012

X. Fu
M. Betts
Q. Wang
ZTE
D. McDysan
A. Malis
Verizon
July 4, 2011

RSVP-TE extensions for latency and loss traffic engineering application
draft-fuxh-ccamp-delay-loss-rsvp-te-ext-00

Abstract

The key driver for latency is stock/commodity trading applications. Financial or trading companies are very focused on end-to-end private pipe line latency optimizations that improve things 2-3 ms. Latency and latency SLA is one of the key parameters that these "high value" customers use to select a private pipe line provider. This document extends RSVP-TE protocol to promote SLA experince of latency application.

Status of this Memo

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

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

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

This Internet-Draft will expire on January 5, 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	3
1.1. Conventions Used in This Document	3
2. SLA Parameters Conveying	3
2.1. Signaling Extensions	4
2.1.1. Latency SLA Parameters subobject	5
2.1.2. Signaling Procedure	7
3. Performance Accumulation and Verification	8
3.1. Signaling Extensions	8
3.1.1. Latency Accumulation Object	8
3.1.1.1. Latency Accumulation sub-TLV	9
3.1.2. Required Latency Object	10
3.1.3. Signaling Procedures	11
4. Security Considerations	12
5. IANA Considerations	13
6. References	13
6.1. Normative References	13
6.2. Informative References	13
Authors' Addresses	13

1. Introduction

End-to-end service optimization based on latency is a key requirement for service provider. It needs to communicate latency of links and nodes including latency and latency variation as a traffic engineering performance metric is a very important requirement. [LATENCY-REQ] describes the requirement of latency traffic engineering application.

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

One end-to-end LSP may be across some Composite Links [CL-REQ]. Even if the transport technology (e.g., OTN) implementing the component links is identical, the latency characteristics of the component links may differ. RSVP-TE message needs to carry a indication for the selection of component links based on the latecny constraint. When one end-to-end LSP traverse a server layer, there will be some latency constraint requirement for the segment route in server layer. RSVP-TE message also needs to carry a indication for the FA selection or FA-LSP creation. This document extends RSVP-TE to indicate that a component links, FA or FA-LSP should meet the minimum and maximum latency value or maximum acceptable latency variation value.

1.1. Conventions Used in This Document

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

2. SLA Parameters Conveying

In order to assign the LSP to one of component links with different latency characteristics, RSVP-TE message MUST convey latency SLA parameter to the end points of Composite Links where it can select one of component links or trigger the creation of lower layer connection which MUST meet latency SLA parameter.

- o The RSVP-TE message needs to carry a indication of request minimum latency, maximum acceptable latency value and maximum acceptable delay variation value for the component link selection or creation. The composite link will take these parameters into account when assigning traffic of LSP to a component link.

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

- o If the FA-LSP is able to form a routing adjacency and/or as a TE link in the client network, the latency value of the FA-LSP can be as an input to a transformation that results in a FA traffic engineering metric and advertised into the client layer routing instances. Note that this metric will include the latency of the links and nodes that the trail traverses.
- o If the latency information of the FA-LSP changes (e.g., due to a maintenance action or failure in OTN rings), the boundary node of the FA-LSP will receive the TE link information advertisement including the latency value which is already changed and if it is over than the threshold and a limit on rate of change, then it will compute the total latency value of the FA-LSP again. If the total latency value of FA-LSP changes, the client layer MUST also be notified about the latest value of FA. The client layer can then decide if it will accept the increased latency or request a new path that meets the latency requirement.
- o When one end-to-end LSP traverse a server layer, there will be some latency constraint requirement for the segment route in server layer. So RSVP-TE message needs to carry a indication of request minimum latency, maximum acceptable latency value and maximum acceptable delay variation value for the FA selection or FA-LSP creation. The boundary nodes of FA-LSP will take these parameters into account for FA selection or FA-LSP creation.

2.1. Signaling Extensions

This document defines extensions to and describes the use of RSVP-TE [RFC3209], [RFC3471], [RFC3473] to explicitly convey the latency SLA parameter for the selection or creation of component link or FA/FA-LSP. Specifically, in this document, Latency SLA Parameters TLV are defined and added into ERO as a subobject.

2.1.1.1. Latency SLA Parameters subobject

A new OPTIONAL subobject of the EXPLICIT_ROUTE Object (ERO) is used to specify the latency SLA parameters including a indication of request minimum latency, request maximum acceptable latency value and request maximum acceptable latency variation value. It can be used for the following scenarios.

- o One end-to-end LSP may traverse a server layer FA-LSP. This subobject of ERO can indicate that FA selection or FA-LSP creation shall be based on this latency constraint. The boundary nodes of multi-layer will take these parameters into account for FA selection or FA-LSP creation.
- o One end-to-end LSP may be across some Composite Links [CL-REQ]. This subobject of ERO can indicate that a traffic flow shall select a component link with some latency constraint values as specified in this subobject.

This Latency SLA Parameters ERO subobject has the following format. It follows a subobject containing the IP address, or the link identifier [RFC3477], associated with the TE link on which it is to be used.

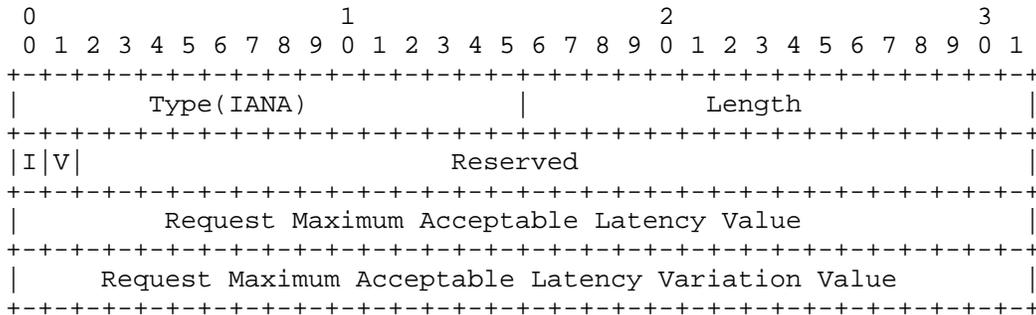


Figure 1: Format of Latency SLA Parameters TLV

- o I bit: a one bit field indicates whether a traffic flow shall select a component link with the minimum latency value or not. It can also indicate whether one end-to-end LSP shall select a FA or trigger a FA-LSP creation with the minimum latency value or not when it traverse a server layer.
- o V bit: a one bit field indicates whether a traffic flow shall select a component link with the minimum latency variation value or not. It can also indicate whether one end-to-end LSP shall select a FA or trigger a FA-LSP creation with the minimum latency

variation value or not when it traverse a server layer.

- o Request Maximum Acceptable Latency Value: a value indicates that a traffic flow shall select a component link with a maximum acceptable latency value. It can also indicate one end-to-end LSP shall select a FA or trigger a FA-LSP creation with a maximum acceptable latency value when it traverse a server layer. It MUST be quantified in units of microseconds and encoded as an integer value.
- o Request Maximum Acceptable Latency Variation Value: a value indicates that a traffic flow shall select a component link with a maximum acceptable latency variation value. It can also indicate one end-to-end LSP shall select a FA or trigger a FA-LSP creation with a maximum acceptable latency variation value when it traverse a server layer. It MUST be quantified in units of nanosecond and encoded as an integer value.

Following is an example about how to use these parameters. Assume there are following component links within one composite link.

- o Component link1: latency = 50 ms, latency variation = 15 ns
- o Component link2: latency = 100 ms, latency variation = 6 ns
- o Component link3: latency = 200 ms, latency variation = 3 ns
- o Component link4: latency = 300 ms, latency variation = 1 ns

Assume there are following request information.

- o Request minimum latency = FALSE
- o Request minimum latency variation= FALSE
- o Maximum Acceptable Latency Value= 150 ms
- o Maximum Acceptable Latency Variation Value = 10 ns

Only Component link2 could be qualified.

- o Request minimum latency = FALSE
- o Request minimum latency variation= FALSE

- o Maximum Acceptable Latency Value= 350 ms
- o Maximum Acceptable Latency Variation Value = 10 ns

Component link2/3/4 could be qualified. Which component link is selected depends on local policy.

- o Request minimum latency = FALSE
- o Request minimum latency variation= TRUE
- o Maximum Acceptable Latency Value= 350 ms
- o Maximum Acceptable Latency Variation Value = 10 ns

Only Component link4 could be qualified.

- o Request minimum latency = TRUE
- o Request minimum latency variation= FALSE
- o Maximum Acceptable Latency Value= 350 ms
- o Maximum Acceptable Latency Variation Value = 10 ns

Only Component link2 could be qualified.

Request minimum latency = TRUE

Request minimum latency variation= TRUE

Maximum Acceptable Latency Value= 350 ms

Maximum Acceptable Latency Variation Value = 10 ns

In this case, there is no any qualified component links. But priority may be used for latency and variation, so one of component links could be still selected.

2.1.2. Signaling Procedure

When a intermediate node receives a PATH message containing ERO and finds that there is a Latency SLA Parameters ERO subobject immediately behind the IP address or link address sub-object related to itself, if the node determines that it's a region edge node of FA-LSP or an end point of a composite link [CL-REQ], then, this node

extracts latency SLA parameters (i.e., request minimum, request maximum acceptable and request maximum acceptable latency variation value) from Latency SLA Parameters ERO subobject. This node used these latency parameters for FA selection, FA-LSP creation or component link selection. If the intermediate node couldn't support the latency SLA, it MUST generate a PathErr message with a "Latency SLA unsupported" indication (TBD by INNA). If the intermediate node couldn't select a FA or component link, or create a FA-LSP which meet the latency constraint defined in Latency SLA Parameters ERO subobject, it must generate a PathErr message with a "Latency SLA parameters couldn't be met" indication (TBD by INNA).

3. Performance Accumulation and Verification

Latency accumulation and verification applies where the full path of an multi-domain (e.g., Inter-AS, Inter-Area or Multi-Layer) TE LSP can't be or is not determined at the ingress node of the TE LSP. This is most likely to arise owing to TE visibility limitations. If all domains support to communicate latency as a traffic engineering metric parameter, one end-to-end optimized path with delay constraint (e.g., less than 10 ms) which satisfies latency SLAs parameter could be computed by BRPC [RFC5441] in PCE. Otherwise, it could use the mechanism defined in this section to accumulat the latency of each links and nodes along the path which is across multi-domain.

Latency accumulation and verification also applies where not all domains could support the communication latency as a traffic engineering metric parameter. The required latency could be signaled by RSVP-TE (i.e., Path and Resv message). Intermediate nodes could reject the request (Path or Resv message) if the accumulated latency is not achievable. This is essential in multiple AS use cases, but may not be needed in a single IGP level/area if the IGP is extended to convey latency information.

One domain may need to know that other domains support latency accumulation. It could be discovered in some automatic way. PCEs in different domains may play a role here. It is for further study.

3.1. Signaling Extensions

3.1.1. Latency Accumulation Object

An Latency Accumulation Object is defined in this document to support the accumulation and verification of the latency. This object which can be carried in a Path/Resv message may includes two sub-TLVs. Latency Accumulation Object has the following format.

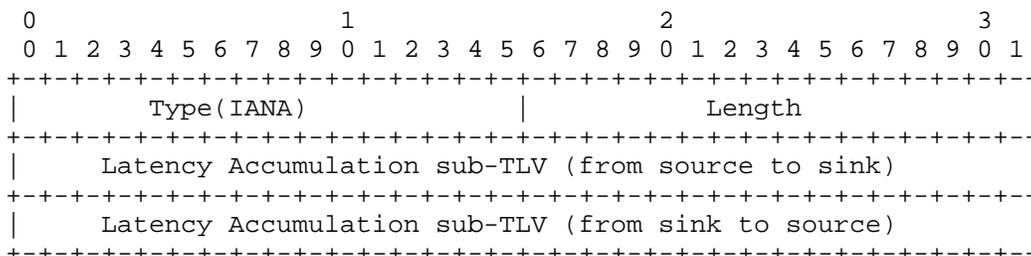


Figure 2: Format of Accumulated Latency Object

- o Latency Accumulation sub-TLV (from source to sink): It is used to accumulate the latency from source to sink along the unidirectional or bidirectional LSP. A Path message for unidirectional and bidirectional LSP must includes this sub-TLV. When sink node receives the Path message including this sub-TLV, it must copy this sub-TLV into Resv message. So the source node can receive the latency accumulated value (i.e., sum) from itself to sink node which can be used for latency verification.
- o Latency Accumulation sub-TLV (from sink to source): It is used to accumulate the latency from sink to source along the bidirectional LSP. A Resv message for the bidirectional LSP must includes this sub-TLV. So the source node can get the latency accumulated value (i.e., sum) of round-trip which can be used for latency verification. It MUST be quantified in units of microseconds and encoded as an integer value.

3.1.1.1. Latency Accumulation sub-TLV

The Sub-TLV format is defined in the next picture.

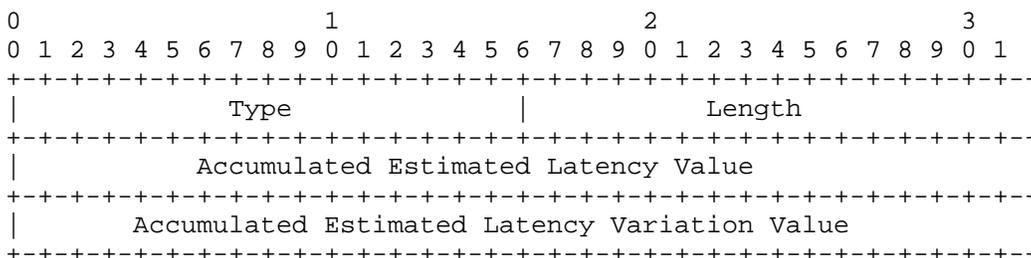


Figure 3: Format of Latency Accumulation sub-TLV

- o Type: sub-TLV type

- * 0: It indicates the sub-TLV is for the latency accumulation from source to sink node along the LSP.
- * 1: It indicates the sub-TLV is for the latency accumulation from sink to source node along the LSP.
- o Length: length of the sub-TLV value in bytes.
- o Accumulated Estimated Latency Value: a value indicates the sum of each links and nodes' latency along one direction of LSP. It MUST be quantified in units of microseconds and encoded as an integer value.
- o Accumulated Estimated Latency Variation Value: a value indicates the sum of each links and nodes' latency variation along one direction of LSP. Since latency variation is accumulated non-linearly. Latency variation accumulation should be in a lower priority. It MUST be quantified in units of nanosecond and encoded as an integer value.

3.1.2. Required Latency Object

A required latency could be signaled by RSVP-TE message (i.e., Path and Resv). This object is carried in the LSP_ATTRIBUTES object of Path/Resv message, object that is defined in [RFC5420]. Intermediate nodes could reject the request (Path or Resv message) if the accumulated latency exceeds require latency value in the Required Latency Object.

If the accumulated latency is not achievable, there is no necessary to accumulate the latency for remaining domain or nodes. In order to balance the load across network links more efficiently if the absolute minimum latency is not required, intermediate nodes could choose a cost-effective path if the requested latency could easily be met. Note that this would apply inter-AS if the IGP is extended to advertise latency.

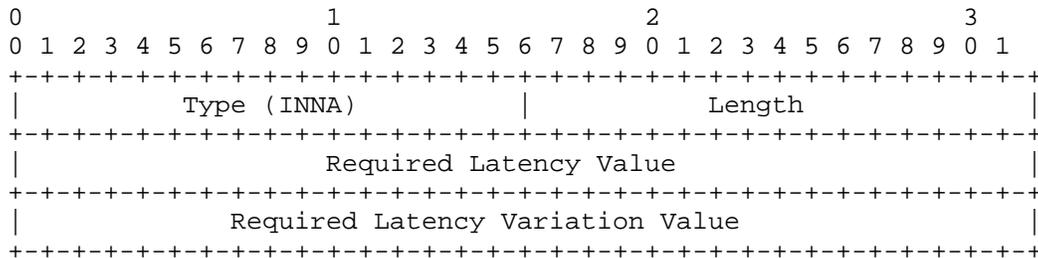


Figure 4: Required Latency Object

- o Required Latency Value: The accumulated estimated latency value should not exceed this value. It MUST be quantified in units of microseconds and encoded as an integer value.
- o Required Latency Variation Value: The accumulated estimated latency variation value should not exceed this value. It MUST be quantified in units of microseconds and encoded as an integer value.

3.1.3. Signaling Procedures

When the source node desires to accumulate (i.e., sum) the total latency of one end-to-end LSP, the "Latency Accumulating desired" flag (value TBD) should be set in the LSP_ATTRIBUTES object of Path/Resv message, object that is defined in [RFC5420]. If the source node makes the intermediate node have the capability to verify the accumulated latency, the "Latency Verifying desired" flag (value TBD) should be also set in the LSP_ATTRIBUTES object of Path/Resv message.

A source node initiates latency accumulation for a given LSP by adding Latency Accumulation object to the Path message. The Latency Accumulation object only includes one sub-TLV (sub-TLV type=0) where it is going to accumulate the latency value of each links and nodes along path from source to sink. If latency verifying is desired, the source node also adds the Required Latency Object to the Path message.

When the downstream node receives Path message and if the "Latency Accumulating desired" is set in the LSP_ATTRIBUTES, it accumulates the latency of link and node based on the accumulated latency value of the sub-TLV (sub-TLV type=0) in Latency Accumulation object before it sends Path message to downstream.

If the "Latency Verifying desired" is set in the LSP_ATTRIBUTES, downstream node will check whether the Accumulated Estimated Latency and Variation value exceeds the Required Latency and Variation value. If the accumulated latency is not achievable, there is no necessary to accumulate the latency for remaining domain or nodes. It MUST generate a error message with a "Accumulated Latency couldn't meet the required latency" indication (TBD by INNA).

If the intermediate node (e.g., entry node of one domain) couldn't support the latency accumulation function, it MUST generate a error message with a "Latency Accumulation unsupported" indication (TBD by INNA).

If the intermediate node (e.g., entry node of one domain) couldn't support the latency verify function, it MUST generate a error message

with a "Latency Verify unsupported" indication (TBD by INNA).

When the sink node of LSP receives the Path message and the "Latency Accumulating desired" is set in the LSP_ATTRIBUTES, it copy the Accumulated Estimated Latency and Variation value in the Latency Accumulation sub-TLV (sub-TLV type=0) of Path message into the one of Resv message which will be forwarded hop by hop in the upstream direction until it arrives the source node. Then source node can get the latency sum value from source to sink for unidirectional and bidirectional LSP.

If the LSP is a bidirectional one and the "Latency Accumulating desired" is set in the LSP_ATTRIBUTES, it adds another Latency Accumulation sub-TLV (sub-TLV type=1) into the Latency Accumulation object of Resv message where latency of each links and nodes along path will be accumulated from sink to source into this sub-TLV.

If the LSP is a bidirectional one and the "Latency Verifying desired" is set in the LSP_ATTRIBUTES, it copy the Required Latency and Variation value in the Required Latency Object of Path message into the one of Resv message.

When the upstream node receives Resv message and if the "Latency Accumulating desired" is set in the LSP_ATTRIBUTES, it accumulates the latency of link and node based on the latency value in sub-TLV (sub-TLV type=1) before it continues to sends Resv message.

If the "Latency Verifying desired" is set in the LSP_ATTRIBUTES, it will check whether the latency sum of Accumulated Estimated Latency and Variation value in each Latency Accumulation sub-TLV exceeds the Required Latency and Variation value. If the accumulated latency is not achievable, there is no necessary to accumulate the latency for remaining domain or nodes. It MUST generate a error message with a "Accumulated Latency couldn't meet the required latency" indication (TBD by INNA).

After source node receive Resv message, it can get the total latency value of one way or round-trip from Latency Accumulation object. So it can confirm whether the latency value meet the latency SLA or not.

4. Security Considerations

TBD

5. IANA Considerations

TBD

6. References

6.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC3209] Awduche, D., Berger, L., Gan, D., Li, T., Srinivasan, V., and G. Swallow, "RSVP-TE: Extensions to RSVP for LSP Tunnels", RFC 3209, December 2001.
- [RFC3473] Berger, L., "Generalized Multi-Protocol Label Switching (GMPLS) Signaling Resource ReserVation Protocol-Traffic Engineering (RSVP-TE) Extensions", RFC 3473, January 2003.
- [RFC3477] Kompella, K. and Y. Rekhter, "Signalling Unnumbered Links in Resource ReSerVation Protocol - Traffic Engineering (RSVP-TE)", RFC 3477, January 2003.
- [RFC3630] Katz, D., Kompella, K., and D. Yeung, "Traffic Engineering (TE) Extensions to OSPF Version 2", RFC 3630, September 2003.
- [RFC4203] Kompella, K. and Y. Rekhter, "OSPF Extensions in Support of Generalized Multi-Protocol Label Switching (GMPLS)", RFC 4203, October 2005.

6.2. Informative References

- [CL-REQ] C. Villamizar, "Requirements for MPLS Over a Composite Link", draft-ietf-rtgwg-cl-requirement-02 .
- [G.709] ITU-T Recommendation G.709, "Interfaces for the Optical Transport Network (OTN)", December 2009.
- [LATENCY-REQ] X. Fu, "GMPLS extensions to communicate latency as a traffic engineering performance metric", draft-wang-ccamp-latency-te-metric-03 .

Authors' Addresses

Xihua Fu
ZTE

Email: fu.xihua@zte.com.cn

Malcolm Betts
ZTE

Email: malcolm.betts@zte.com.cn

Qilei Wang
ZTE

Email: wang.qilei@zte.com.cn

Dave McDysan
Verizon

Email: dave.mcdysan@verizon.com

Andrew Malis
Verizon

Email: andrew.g.malis@verizon.com

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: January 5, 2012

X. Fu
M. Betts
Q. Wang
ZTE
D. McDysan
A. Malis
Verizon
July 4, 2011

Framework for latency and loss traffic engineering application
draft-fuxh-ccamp-delay-loss-te-framework-00

Abstract

Latency and packet loss is such requirement that must be achieved according to the Service Level Agreement (SLA) / Network Performance Objective (NPO) between customers and service providers. Latency and packet loss can be associated with different service level. The user may select a private line provider based on the ability to meet a latency and loss SLA.

The key driver for latency and loss is stock/commodity trading applications that use data base mirroring. A few milli seconds and packet loss can impact a transaction. Financial or trading companies are very focused on end-to-end private pipe line latency optimizations that improve things 2-3 ms. Latency/loss and associated SLA is one of the key parameters that these "high value" customers use to select a private pipe line provider. Other key applications like video gaming, conferencing and storage area networks require stringent latency, loss and bandwidth.

This document describes requirements to communicate latency and packet loss as a traffic engineering performance metric in today's network which is consisting of potentially multiple layers of packet transport network and optical transport network in order to meet the latency/loss SLA between service provider and his customers.

Status of this Memo

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

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

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

This Internet-Draft will expire on January 5, 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 4
 - 1.1. Conventions Used in This Document 4
- 2. Latency and Loss Report 4
- 3. Requirements Identification 5
- 4. Control Plane Implication 7
- 5. Security Considerations 9
- 6. IANA Considerations 9
- 7. References 9
 - 7.1. Normative References 9
 - 7.2. Informative References 10
- Authors' Addresses 10

1. Introduction

Current operation and maintenance mode of latency and packet loss measurement is high in cost and low in efficiency. The latency and packet loss can only be measured after the connection has been established, if the measurement indicates that the latency SLA is not met then another path is computed, setup and measured. This "trial and error" process is very inefficient. To avoid this problem a means of making an accurate prediction of latency and packet loss before a path is established is required.

This document describes the requirements and control plane implication to communicate latency and packet loss as a traffic engineering performance metric in today's network which is consisting of potentially multiple layers of packet transport network and optical transport network in order to meet the latency and packet loss SLA between service provider and his customers.

1.1. Conventions Used in This Document

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

2. Latency and Loss Report

This section isn't going to say how latency or packet loss is measured. How to measure has been provided in ITU-T [Y.1731], [G.709] and [ietf-mpls-loss-delay]. It's purpose is to define what it is sufficiently clear that mechanisms could be defined to measure it, and so that independent implementations will report the same thing. If control plane wish to define the ability to report latency and packet loss, control plane must be clear what it are reporting.

Packet/Frame loss probability is expressed as a percentage of the number of service packets/frames not delivered divided by the total number of service frames during time interval T. Loss is always measured by sending a measurement packet or frame from measurement point to its reception and reception sending back a response.

The link of latency is the time interval between the propagation of an electrical signal and its reception. Latency is always measured by sending a measurement packet or frame from measurement point to its reception. In some usages, latency is measured by sending a packet/frame that is returned to the sender and the round-trip time is considered the latency of bidirectional co-routed or associated LSP. One way time is considered as the latency of unidirectional

LSP. The one way latency may not be half of the round-trip latency in the case that the transmit and receive directions of the path are of unequal lengths.

Control plane should report two components of the delay, "static" and "dynamic". The dynamic component is caused by traffic loading. What is reporting for "dynamic" portion is approximation.

Latency on a connection has two sources: Node latency which is caused by the node as a result of process time in each node and: Link latency as a result of packet/frame transit time between two neighbouring nodes or a FA-LSP/Composit Link [CL-REQ]. The average latency of node should be reported. It is simpler to add node latency to the link delay vs. carrying a separate parameter and does not hide any important information. Latency variation is a parameter that is used to indicate the variation range of the latency value. Latency, latecny variation value must be reported as a average value which is calculated by data plane.

3. Requirements Identification

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

- o REQ #1: The solution MUST provide a means to communicate latency, latency variation and packet loss of links and nodes as a traffic engineering performance metric into IGP.
- o REQ #2: Latency, latency variation and packet loss may be unstable, for example, if queueing latency were included, then IGP could become unstable. The solution MUST provide a means to control latency and loss IGP message advertisement and avoid unstable when the latency, latency variation and packet loss value changes.
- o REQ #3: Path computation entity MUST have the capability to compute one end-to-end path with latency and packet loss constraint. for example, it has the capability to compute a route with X amount bandwidth with less than Y ms of latency and Z% packet loss limit based on the latency and packet loss traffic engineering database. It MUST also support the path computation with routing constraints combination with pre-defined priorities,

e.g., SRLG diversity, latency, loss and cost.

- o REQ #4: One end-to-end LSP may traverse some Composite Links [CL-REQ]. Even if the transport technology (e.g., OTN) implementing the component links is identical, the latency and packet loss characteristics of the component links may differ. In order to assign the LSP to one of component links with different latency and packet loss characteristics, the solution SHOULD provide a means to indicate that a traffic flow should select a component link with minimum latency and/or packet loss, maximum acceptable latency and/or packet loss value and maximum acceptable delay variation value as specified by protocol. The endpoints of Composite Link will take these parameters into account for component link selection or creation.
- o REQ #5: One end-to-end LSP may traverse a server layer. There will be some latency and packet loss constraint requirement for the segment route in server layer. The solution SHALL provide a means to indicate FA selection or FA-LSP creation with minimum latency and/or packet loss, maximum acceptable latency and/or packet loss value and maximum acceptable delay variation value. The boundary nodes of FA-LSP will take these parameters into account for FA selection or FA-LSP creation.
- o REQ #6: The solution SHOULD provide a means to accumulate (e.g., sum) of latency information of links and nodes along one LSP across multi-domain (e.g., Inter-AS, Inter-Area or Multi-Layer) so that a latency validation decision can be made at the source node. One-way and round-trip latency collection along the LSP by signaling protocol and latency verification at the end of LSP should be supported. The accumulation of the delay is "simple" for the static component i.e. its a linear addition, the dynamic/network loading component is more interesting and would involve some estimate of the "worst case". However, method of deriving this worst case appears to be more in the scope of Network Operator policy than standards i.e. the operator needs to decide, based on the SLAs offered, the required confidence level.
- o REQ #7: Some customers may insist on having the ability to re-route if the latency and loss SLA is not being met. If a "provisioned" end-to-end LSP latency and/or loss could not meet the latency and loss agreement between operator and his user, The solution SHOULD support pre-defined or dynamic re-routing to handle this case based on the local policy. The latency performance of pre-defined protection or dynamic re-routing LSP MUST meet the latency SLA parameter.

- o REQ #8: If a "provisioned" end-to-end LSP latency and/or loss performance is improved because of some segment performance promotion, the solution SHOULD support the re-routing to optimize latency and/or loss end-to-end cost.
- o REQ #9: As a result of the change of latency and loss in the LSP, current LSP may be frequently switched to a new LSP with a appropriate latency and packet loss value. In order to avoid this, the solution SHOULD indicate the switchover of the LSP according to maximum acceptable change latency and packet loss value.

4. Control Plane Implication

- o The latency and packet loss performance metric MUST be advertised into path computation entity by IGP (etc., OSPF-TE or IS-IS-TE) to perform route computation and network planning based on latency and packet loss SLA target. Latency, latency variation and packet loss value MUST be reported as a average value which is calculated by data plane. Latency and packet loss characteristics of these links and nodes may change dynamically. In order to control IGP messaging and avoid being unstable when the latency, latency variation and packet loss value changes, a threshold and a limit on rate of change MUST be configured to control plane. If any latency and packet loss values change and over than the threshold and a limit on rate of change, then the change MUST be notified to the IGP again.
- o Link latency attribute may also take into account the latency of a network element (node), i.e., the latency between the incoming port and the outgoing port of a network element. If the link attribute is to include node latency AND link latency, then when the latency calculation is done for paths traversing links on the same node then the node latency can be subtracted out.
- o When the Composite Links [CL-REQ] is advertised into IGP, there are following considerations.
 - * The latency and packet loss of composite link may be the range (e.g., at least minimum and maximum) latency value of all component links. It may also be the maximum latency value of all component links. In these cases, only partial information is transmitted in the IGP. So the path computation entity has insufficient information to determine whether a particular path can support its latency and packet loss requirements. This leads to signaling crankback. So IGP may be extended to advertise latency and packet of each component link within one

Composite Link having an IGP adjacency.

- o One end-to-end LSP (e.g., in IP/MPLS or MPLS-TP network) may traverse a FA-LSP of server layer (e.g., OTN rings). The boundary nodes of the FA-LSP SHOULD be aware of the latency and packet loss information of this FA-LSP.
 - * If the FA-LSP is able to form a routing adjacency and/or as a TE link in the client network, the total latency and packet loss value of the FA-LSP can be as an input to a transformation that results in a FA traffic engineering metric and advertised into the client layer routing instances. Note that this metric will include the latency and packet loss of the links and nodes that the trail traverses.
 - * If total latency and packet loss information of the FA-LSP changes (e.g., due to a maintenance action or failure in OTN rings), the boundary node of the FA-LSP will receive the TE link information advertisement including the latency and packet value which is already changed and if it is over than the threshold and a limit on rate of change, then it will compute the total latency and packet value of the FA-LSP again. If the total latency and packet loss value of FA-LSP changes, the client layer MUST also be notified about the latest value of FA. The client layer can then decide if it will accept the increased latency and packet loss or request a new path that meets the latency and packet loss requirement.
- o Restoration, protection and equipment variations can impact "provisioned" latency and packet loss (e.g., latency and packet loss increase). The change of one end-to-end LSP latency and packet loss performance MUST be known by source and/or sink node. So it can inform the higher layer network of a latency and packet loss change. The latency or packet loss change of links and nodes will affect one end-to-end LSP's total amount of latency or packet loss. Applications can fail beyond an application-specific threshold. Some remedy mechanism could be used.
 - * Pre-defined protection or dynamic re-routing could be triggered to handle this case. In the case of predefined protection, large amounts of redundant capacity may have a significant negative impact on the overall network cost. Service provider may have many layers of pre-defined restoration for this transfer, but they have to duplicate restoration resources at significant cost. Solution should provides some mechanisms to avoid the duplicate restoration and reduce the network cost. Dynamic re-routing also has to face the risk of resource limitation. So the choice of mechanism MUST be based on SLA or

policy. In the case where the latency SLA can not be met after a re-route is attempted, control plane should report an alarm to management plane. It could also try restoration for several times which could be configured.

5. Security Considerations

The use of control plane protocols for signaling, routing, and path computation of latency and loss opens security threats through attacks on those protocols. The control plane may be secured using the mechanisms defined for the protocols discussed. For further details of the specific security measures refer to the documents that define the protocols ([RFC3473], [RFC4203], [RFC4205], [RFC4204], and [RFC5440]). [GMPLS-SEC] provides an overview of security vulnerabilities and protection mechanisms for the GMPLS control plane.

6. IANA Considerations

This document makes not requests for IANA action.

7. References

7.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC3209] Awduche, D., Berger, L., Gan, D., Li, T., Srinivasan, V., and G. Swallow, "RSVP-TE: Extensions to RSVP for LSP Tunnels", RFC 3209, December 2001.
- [RFC3473] Berger, L., "Generalized Multi-Protocol Label Switching (GMPLS) Signaling Resource ReserVation Protocol-Traffic Engineering (RSVP-TE) Extensions", RFC 3473, January 2003.
- [RFC3477] Kompella, K. and Y. Rekhter, "Signalling Unnumbered Links in Resource ReSerVation Protocol - Traffic Engineering (RSVP-TE)", RFC 3477, January 2003.
- [RFC3630] Katz, D., Kompella, K., and D. Yeung, "Traffic Engineering (TE) Extensions to OSPF Version 2", RFC 3630, September 2003.
- [RFC4203] Kompella, K. and Y. Rekhter, "OSPF Extensions in Support

of Generalized Multi-Protocol Label Switching (GMPLS)",
RFC 4203, October 2005.

7.2. Informative References

- [CL-REQ] C. Villamizar, "Requirements for MPLS Over a Composite Link", draft-ietf-rtgwg-cl-requirement-02 .
- [G.709] ITU-T Recommendation G.709, "Interfaces for the Optical Transport Network (OTN)", December 2009.
- [Y.1731] ITU-T Recommendation Y.1731, "OAM functions and mechanisms for Ethernet based networks", Feb 2008.
- [ietf-mpls-loss-delay]
D. Frost, "Packet Loss and Delay Measurement for MPLS Networks", draft-ietf-mpls-loss-delay-03 .

Authors' Addresses

Xihua Fu
ZTE

Email: fu.xihua@zte.com.cn

Malcolm Betts
ZTE

Email: malcolm.betts@zte.com.cn

Qilei Wang
ZTE

Email: wang.qilei@zte.com.cn

Dave McDysan
Verizon

Email: dave.mcdysan@verizon.com

Andrew Malis
Verizon

Email: andrew.g.malis@verizon.com

MPLS Working Group
Internet-Draft
Intended status: Standards Track
Expires: December 3, 2011

F. Zhang, Ed.
ZTE
R. Jing
China Telecom
June 01, 2011

RSVP-TE Extensions to Establish Associated Bidirectional LSP
draft-ietf-ccamp-mpls-tp-rsvpte-ext-associated-lsp-01

Abstract

The MPLS Transport Profile (MPLS-TP) requirements document [RFC5654], describes that MPLS-TP MUST support associated bidirectional point-to-point LSPs.

This document provides a method to bind two unidirectional Label Switched Paths (LSPs) into an associated bidirectional LSP. The association is achieved by using a new Association Type in the Extended ASSOCIATION object.

Status of this Memo

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

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

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

This Internet-Draft will expire on December 3, 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	3
2. Conventions used in this document	4
3. Association of Two Reverse Unidirectional LSPs	4
3.1. Provisioning Model	4
3.2. Signaling Procedure	4
3.2.1. Single Sided Provisioning Model	5
3.2.2. Double Sided Provisioning Model	6
3.2.3. Asymmetric Bandwidth LSPs	8
3.2.3.1. Error Handling	9
3.3. Recovery Considerations	10
4. Extensions to the Extended ASSOCIATION object	10
5. REVERSE_TSPEC Object	13
6. IANA Considerations	13
6.1. Association Type	13
6.2. REVERSE_TSPEC Object	14
7. Security Considerations	14
8. Acknowledgement	14
9. References	15
9.1. Normative references	15
9.2. Informative References	15
Authors' Addresses	16

1. Introduction

The MPLS Transport Profile (MPLS-TP) requirements document [RFC5654] describes that MPLS-TP MUST support associated bidirectional point-to-point LSPs. Furthermore, an associated bidirectional LSP is useful for protection switching, for Operations, Administrations and Maintenance (OAM) messages that require a reply path.

The requirements described in [RFC5654] are specifically mentioned in Section 2.1. (General Requirements), and are repeated below:

7. MPLS-TP MUST support associated bidirectional point-to-point LSPs.

11. The end points of an associated bidirectional LSP MUST be aware of the pairing relationship of the forward and reverse LSPs used to support the bidirectional service.

12. Nodes on the LSP of an associated bidirectional LSP where both the forward and backward directions transit the same node in the same (sub)layer as the LSP SHOULD be aware of the pairing relationship of the forward and the backward directions of the LSP.

14. MPLS-TP MUST support bidirectional LSPs with asymmetric bandwidth requirements, i.e., the amount of reserved bandwidth differs between the forward and backward directions.

50. The MPLS-TP control plane MUST support establishing associated bidirectional P2P LSP including configuration of protection functions and any associated maintenance functions.

The above requirements are also repeated in [I-D.ietf-ccamp-mpls-tp-cp-framework].

The notion of association, as well as the corresponding Resource reSerVation Protocol (RSVP) ASSOCIATION object, is defined in [RFC4872], [RFC4873] and [I-D.ietf-ccamp-assoc-info]. In that context, the object is used to associate recovery LSPs with the LSP they are protecting. This object also has broader applicability as a mechanism to associate RSVP state, and [I-D.ietf-ccamp-assoc-ext] defines the Extended ASSOCIATION object that can be more generally applied.

This document provides a method to bind two unidirectional Label Switched Paths (LSPs) into an associated bidirectional LSP. The association is achieved by using a new Association Type in the Extended ASSOCIATION object.

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. Association of Two Reverse Unidirectional LSPs

3.1. Provisioning Model

The associated bidirectional LSP's forward and backward directions are set up, monitored, and protected independently as required by [RFC5654]. Configuration information regarding the LSPs can be sent to one end or both ends of the LSP. Depending on the method chosen, there are two models of signaling associated bidirectional LSP. The first model is the single sided provisioning, the second model is the double sided provisioning.

For the single sided provisioning, the configurations are sent to one end. Firstly, a unidirectional tunnel is configured on this end, then a LSP under this tunnel is initiated with the Extended ASSOCIATION object carried in the Path message to trigger the peer end to set up the corresponding reverse TE tunnel and LSP.

For the double sided provisioning, the two unidirectional TE tunnels are configured independently, then the LSPs under the tunnels are signaled with the Extended ASSOCIATION objects carried in the Path message to indicate each other to associate the two LSPs together to be an associated bidirectional LSP.

A number of scenarios exist for binding LSPs together to be an associated bidirectional LSP. These include: (1) both of them do not exist; (2) both of them exist; (3) one LSP exists, but the other one need to be established. In all scenarios described, the provisioning models discussed above are applicable.

3.2. Signaling Procedure

This section describes the signaling procedures for associating bidirectional LSPs.

Consider the topology described in Figure 1. (An example of associated bidirectional LSP). The LSP1 [via nodes A,D,B] (from west to east) and LSP2 [via nodes B,D,C,A] (from east to west) are being established or have been established. These LSPs can be bound together to form an associated bidirectional LSP.

LSP1 is uniquely identified [I-D.ietf-mpls-tp-identifiers] by: West-Global_ID::West-Node_ID::West-Tunnel_Num::West-LSP_Num.

LSP2 is uniquely identified [I-D.ietf-mpls-tp-identifiers] by: East-Global_ID::East-Node_ID::East-Tunnel_Num::East-LSP_Num.

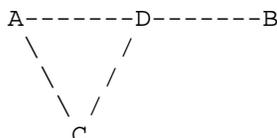


Figure 1: An example of associated bidirectional LSP

3.2.1. Single Sided Provisioning Model

For the single sided provisioning model, LSP1 is triggered by LSP2 or LSP2 is triggered by LSP1. When LSP2 is triggered by LSP1, according to the scenarios described above, the following cases may occur:

1. Both LSPs do not exist.

LSP1 is initialized at node A with the Extended ASSOCIATION object inserted in the Path message, Association Type is set to "Association of two reverse unidirectional LSPs", Association ID set to West-LSP_Num, Association Source set to West-Node_ID, Global Association Source set to West-Global_ID, and Extended Association ID set to West-Tunnel_Num. Terminating node B is triggered to set up LSP2 by the received Extended ASSOCIATION object with the Association Type set to the value "Association of two reverse unidirectional LSPs", the Association Object inserted in LSP2's Path message is the same as in LSP1's Path message.

2. LSP1 exists, LSP2 needs to be established.

LSP1 is refreshed with the Extended ASSOCIATION object inserted in the Path message, Association Type is set to "Association of two reverse unidirectional LSPs", Association ID set to West-LSP_Num, Association Source set to West-Node_ID, Global Association Source set to West-Global_ID, and Extended Association ID set to West-Tunnel_Num. Terminating node B is triggered to set up LSP2 by the received Extended ASSOCIATION object with the Association Type set to the value "Association of two reverse unidirectional LSPs", the Association Object inserted in LSP2's Path message is the same as in LSP1's Path message.

3. LSP1 does not exist, LSP2 has been established.
LSP1 is initialized with the Extended ASSOCIATION object inserted in the Path message, Association Type is set to "Association of two reverse unidirectional LSPs", Association ID set to East-LSP_Num, Association Source set to East-Node_ID, Global Association Source set to East-Global_ID, and Extended Association ID set to East-Tunnel_Num. Terminating node B is triggered to refresh LSP2's Path message, with the received Extended ASSOCIATION object inserted.

4. Both LSP1 and LSP2 exist.
LSP1 is refreshed with the Extended ASSOCIATION object inserted in the Path message, Association Type is set to "Association of two reverse unidirectional LSPs", Association ID set to East-LSP_Num, Association Source set to East-Node_ID, Global Association Source set to East-Global_ID, and Extended Association ID set to East-Tunnel_Num. Terminating node B is triggered to refresh LSP2's Path message, with the received Extended ASSOCIATION object inserted.

When LSP1 is triggered by LSP2, the same rules are applicable. Based on the same values of the Association objects in the two LSPs' Path message, the two LSPs can be bound together to be an associated bidirectional LSP.

3.2.2. Double Sided Provisioning Model

For the double sided provisioning model, Similarly, according to the scenarios described above, the following cases may occur:

1. LSP1 and LSP2 do not exist.
LSP1 and LSP2 are concurrently initialized with the Extended ASSOCIATION object inserted in the their Path messages, For LSP1, Association Type is set to "Association of two reverse unidirectional LSPs", Association ID set to West-LSP_Num, Association Source set to West-Node_ID, Global Association Source set to West-Global_ID, and Extended Association ID set to West-Tunnel_Num. For LSP2, Association Type is set to "Association of two reverse unidirectional LSPs", Association ID is set to East-LSP_Num, Association Source set to East-Node_ID, Global Association Source set to East-Global_ID, and Extended Association ID set to East-Tunnel_Num. According to the general rules defined in [I-D.ietf-ccamp-assoc-ext], the two LSPs cannot be bound together to be an associated bidirectional LSP because of the different values. In this case, the two edge nodes firstly MUST compare their Global-Node_ID, then the bigger one sends Path refresh message, replacing the old Extended ASSOCIATION object with the new Extended ASSOCIATION object carried in the reverse LSP. Based on this Path refresh message, the two LSPs can be

bounded together to be an associated bidirectional LSP also.

2. LSP1 exists, LSP2 needs to be established.

For LSP1, Association Type is set to "Association of two reverse unidirectional LSPs", Association ID set to West-LSP_Num, Association Source set to West-Node_ID, Global Association Source set to West-Global_ID, and Extended Association ID set to West-Tunnel_Num. For LSP2, Node B has known the existence of LSP1, so the Association Type is set to "Association of two reverse unidirectional LSPs", Association ID set to West-LSP_Num, Association Source set to West-Node_ID, Global Association Source set to West-Global_ID, and Extended Association ID set to West-Tunnel_Num.

3. LSP1 does not exist, LSP2 has been established.

For LSP1, Node A has known the existence of LSP2. So the Association Type is set to "Association of two reverse unidirectional LSPs", Association ID set to East-LSP_Num, Association Source set to East-Node_ID, Global Association Source set to East-Global_ID, and Extended Association ID set to East-Tunnel_Num. For LSP2, Node B does not know the existence of LSP1, so Association Type is set to "Association of two reverse unidirectional LSPs", Association ID set to East-LSP_Num, Association Source set to East-Node_ID, Global Association Source set to East-Global_ID, and Extended Association ID set to East-Tunnel_Num.

4. Both LSP1 and LSP2 exist.

In this case, Both node A and Node B know the existence of the reverse LSPs. The two edge nodes firstly MUST compare their Global-Node_ID, then the bigger one sends Path refresh message, with the reverse LSP's identifier inserted in the Extended ASSOCIATION object, and the smaller one sends Path refresh message, with its own LSP's identifier inserted in the Extended ASSOCIATION object. For example, assuming that the node A has the bigger Global-Node_ID. For LSP1, the Association Type is set to "Association of two reverse unidirectional LSPs", Association ID set to East-LSP_Num, Association Source set to East-Node_ID, Global Association Source set to East-Global_ID, and Extended Association ID set to East-Tunnel_Num. For LSP2, the Association Type is set to "Association of two reverse unidirectional LSPs", Association ID set to West-LSP_Num, Association Source set to East-Node_ID, Global Association Source set to East-Global_ID, and Extended Association ID set to East-Tunnel_Num.

Based on the same values of the Association objects in the two LSPs' Path message, the two LSPs can be bound together to be an associated bidirectional LSP.

3.2.3. Asymmetric Bandwidth LSPs

A variety of applications, such as internet services and the return paths of OAM messages, exist and which MAY have different bandwidth requirements for each direction. Additional [RFC5654] also specifies an asymmetric bandwidth requirement. This requirement is specifically mentioned in Section 2.1. (General Requirements), and is repeated below:

14. MPLS-TP MUST support bidirectional LSPs with asymmetric bandwidth requirements, i.e., the amount of reserved bandwidth differs between the forward and backward directions.

The approach for supporting asymmetric bandwidth co-routed bidirectional LSPs is defined in [I-D.ietf-ccamp-asymm-bw-bidir-lsps-bis], which introduces three new objects named UPSTREAM_FLOWSPEC object, UPSTREAM_TSPEC object and UPSTREAM_ADSPEC object to represent the asymmetric upstream traffic flow. For the asymmetric bandwidth associated bidirectional LSPs, the existing SENDER_TSPEC, ADSPEC, and FLOWSPEC are complemented with the addition of a new REVERSE_TSPEC object, which is used exactly in the same fashion as the old SENDER_TSPEC object.

Consider the topology described in Figure 1 in the context of asymmetric associated bidirectional LSP, the following cases may occur:

1. LSP1 and LSP2 do not exist.

For the single sided provisioning, taking LSP2 triggered by LSP1 as an example. The REVERSE_TSPEC object MUST be carried in the LSP1's Path message together with the Extended ASSOCIATION object whose Association Type is "Association of two reverse unidirectional LSPs". The terminating node B is triggered to set up the reverse LSP2 with the corresponding asymmetric bandwidth, and the REVERSE_TSPEC object is converted to the SENDER_TSPEC object in the Path message.

For the double sided provisioning, the REVERSE_TSPEC object MUST be carried in the two LSPs' Path message together with the Extended ASSOCIATION object whose Association Type is "Association of two reverse unidirectional LSPs". Then the two terminating ends MUST compare the values of the SENDER_TSPEC and REVERSE_TSPEC objects in the two Path messages. If the values match, the end with the bigger Global-Node_ID sends Path refresh message, carrying the Extended ASSOCIATION object of the reverse LSP.

2. LSP1 exists, LSP2 needs to be established.

For the single sided provisioning, taking LSP2 triggered by LSP1 as an example. The REVERSE_TSPEC object MUST be carried in the LSP1's Path refresh message together with the Extended ASSOCIATION object whose Association Type is "Association of two reverse unidirectional LSPs". The terminating node B is triggered to set up the reverse LSP2 with the corresponding asymmetric bandwidth, and the REVERSE_TSPEC object is converted to the SENDER_TSPEC object in the Path message.

For the double sided provisioning, the REVERSE_TSPEC object MUST be carried in the LSP1's Path refresh message with the Extended ASSOCIATION object whose Association Type is "Association of two reverse unidirectional LSPs". There is no need to put the REVERSE_TSPEC object in LSP2's Path message, for the Extended ASSOCIATION object has indicated that LSP2 needs to be bound with LSP1.

3. LSP1 does not exist, LSP2 has been established.

For the single sided provisioning, taking LSP2 triggered by LSP1 as an example. There is no need to put the REVERSE_TSPEC object in LSP1's Path message, for the Extended ASSOCIATION object has indicates that LSP1 needs to be bound with LSP2.

For the double sided provisioning, just the same reason, the REVERSE_TSPEC object only needs to be carried in the LSP2's Path refresh message.

4. Both LSP1 and LSP2 exist.

For the single sided provisioning, taking LSP2 triggered by LSP1 as an example. There is no need to put the REVERSE_TSPEC object in LSP1's Path message also for the Extended ASSOCIATION object has indicates that LSP1 needs to be bound with LSP2.

As to the double sided provisioning, just the same reason, the REVERSE_TSPEC object does not need to be carried in the two LSPs' Path messages.

Based on the same values of the Association objects in the two LSPs' Path message, and the match of the REVERSE_TSPEC and SENDER_TSPEC objects in the two LSPs' Path message (if the REVERSE_TSPEC object exists), the two LSPs can be bound together to be an associated bidirectional LSP.

3.2.3.1. Error Handling

Nodes not supporting the new class number of the REVERSE_TSPEC object SHOULD respond with an "Unknown Object Class".

3.3. Recovery Considerations

Consider the topology described in Figure 1, LSP1 and LSP2 form the associated bidirectional LSP. Under the scenario of recovery, a third LSP (LSP3) MAY be used to protect LSP1. LSP3 can be established before or after the failure occurs, it can share the same TE tunnel with LSP1 or not.

In the case that LSP3 is established after the failure occurs, the Extended ASSOCIATION object with LSP2's identifier SHOULD be inserted in LSP3's Path message since LSP2 has already existed. If LSP1 and LSP2 are associated together by the LSP1's identifier, LSP2's Path message is refreshed, an additional Extended ASSOCIATION object with LSP2's identifier are inserted. If LSP1 and LSP2 are bound together by the LSP2's identifier, there is no need to insert an additional Extended ASSOCIATION object in LSP2's Path message.

In the case that LSP3 is established before the failure occurs. For single sided provisioning, LSP3 is refreshed with the Extended ASSOCIATION object, its values are filled by LSP2's identifier. Then LSP2 is refreshed with this Extended ASSOCIATION object or not, see the description in the above paragraph. For double sided provisioning, if node A has the bigger Global-Node_ID than node B, LSP3 is refreshed with the Extended ASSOCIATION object whose values are filled by LSP2's identifier, and LSP2 is refreshed with this Extended ASSOCIATION object or not, see the description in the above paragraph. If node A has the smaller Global-Node_ID than node B, LSP3 is refreshed with the Extended ASSOCIATION object whose values are filled by LSP3's identifier, and LSP2 is refreshed with this Extended ASSOCIATION object.

4. Extensions to the Extended ASSOCIATION object

The Extended ASSOCIATION object is defined in [I-D.ietf-ccamp-assoc-ext], which enables MPLS-TP required identification.

The Extended IPv4 ASSOCIATION object (Class-Num of the form 11bbbbbb with value = 199, C-Type = TBA) has the format:

In order to bind two reverse unidirectional LSPs to be an associated bidirectional LSP, this document defines a new Association Type:

Value	Type
-----	-----
4 (TBD)	Association of two reverse unidirectional LSPs (A)

If the downstream nodes are not aware of the Association Type, they MUST return a PathErr message with error code/sub-code "LSP Admission Failure/Bad Association Type".

Under the context of this Association Type, any node associating an associated bidirectional LSP MUST insert an ASSOCIATION object with the following setting:

- o Association ID:

The Association ID MUST be set to its own signaled LSP ID (default); if known, it MAY be set to the LSP ID of the associated reverse LSP.

- o Association Source:

The Association source MUST be set to the tunnel sender address of this LSP (default); if known, it May be set to the tunnel sender address of the peer node.

- o Global Association Source:

The format is described in [I-D.ietf-ccamp-assoc-ext].

- o Extended Association ID:

Because the two LSPs (one is from west to east, and the other is from east to west) are in different tunnels, the Association ID is insufficient to uniquely identify association for associated bidirectional LSP. Hence, this document adds specific rules: the first 16-bits MUST be set to its own tunnel ID (default); if known, it May be set to the tunnel ID of the the associated reverse tunnel.

As described in [I-D.ietf-ccamp-assoc-ext], association is always done based on matching Path state or Resv state. Upstream initialized association is represented in Extended ASSOCIATION objects carried in Path message and downstream initialized

association is represented in Extended ASSOCIATION objects carried in Resv messages. The new defined association type in this document is only defined for use in upstream initialized association. Thus it can only appear in Extended ASSOCIATION objects signaled in Path message.

The rules associated with the processing of the Extended ASSOCIATION objects in RSVP message are discussed in [I-D.ietf-ccamp-assoc-ext]. It said that in the absence of Association Type-specific rules for identifying association, the included Extended ASSOCIATION objects MUST be identical. This document adds no specific rules, the association will always operate based on the same Extended ASSOCIATION objects.

5. REVERSE_TSPEC Object

The REVERSE_TSPEC object is used in Path, PathTear, PathErr, and Notify message (via sender descriptor). This includes the definition of class type and format. It's class number is TBD (of the form 0bbbbbbb), and class type and format is the same as the SENDER_TSPEC object.

This object modifies the RSVP message-related formats defined in [RFC2205], [RFC3209] and [RFC3473]. See [RFC5511] for the syntax used by RSVP. The format of the sender description for asymmetric associated bidirectional LSPs is:

```
<sender descriptor> ::= <SENDER_TEMPLATE> <SENDER_TSPEC>
                        [<ADSPEC>]
                        [<RCEORD_ROUTE>]
                        [<SUGGESTED_LABEL>]
                        [<RECOVERY_LABEL>]
                        <REVERSE_TSPEC>
```

6. IANA Considerations

IANA is requested to administer assignment of new values for namespace defined in this document and summarized in this section.

6.1. Association Type

Within the current document, a new Association Type is defined in the Extended ASSOCIATION object.

Value	Type
-----	-----
4 (TBD)	Association of two reverse unidirectional LSPs (A)

6.2. REVERSE_TSPEC Object

A new class named REVERSE_TSPEC has been created in the 0bbbbbbb rang (123,TBD) with the following definition:

Class Types or C-types:

Same values as SENDER_TPSCE object (C-Num 12)

There are no other IANA considerations introduced by this document.

7. Security Considerations

This document introduces a new association type, and except this, there are no security issues about the Extended ASSOCIATION object are introduced here.

Furthermore, this document introduces the REVERSE_TSPEC object for use in GMPLS signaling [RFC3473], which is parallel the existing SENDER_TSPEC object. As such, any vulnerabilities that are due to the use of the old SENDER_TSPEC object now apply here also.

Otherwise, this document introduces no additional security considerations. For a general discussion on MPLS and GMPLS related security issues, see the MPLS/GMPLS security framework [RFC5920].

8. Acknowledgement

The authors would like to thank Lou Berger for his great guidance in this work, George Swallow and Jie Dong for the discussion of recovery, Lamberto Sterling for his valuable comments on the section of asymmetric bandwidths, Daniel King for the review of the document, Attila Takacs for the discussion of the provisioning model. At the same time, the authors would also like to acknowledge the contributions of Bo Wu, Xihua Fu, Lizhong Jin, and Wenjuan He for the initial discussions.

9. References

9.1. Normative references

- [I-D.ietf-ccamp-assoc-ext]
Berger, L., Faucheur, F., and A. Narayanan, "RSVP Association Object Extensions", draft-ietf-ccamp-assoc-ext-00 (work in progress), May 2011.
- [I-D.ietf-mpls-tp-identifiers]
Bocci, M., Swallow, G., and E. Gray, "MPLS-TP Identifiers", draft-ietf-mpls-tp-identifiers-04 (work in progress), March 2011.
- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC4872] Lang, J., Rekhter, Y., and D. Papadimitriou, "RSVP-TE Extensions in Support of End-to-End Generalized Multi-Protocol Label Switching (GMPLS) Recovery", RFC 4872, May 2007.
- [RFC4873] Berger, L., Bryskin, I., Papadimitriou, D., and A. Farrel, "GMPLS Segment Recovery", RFC 4873, May 2007.
- [RFC5654] Niven-Jenkins, B., Brungard, D., Betts, M., Sprecher, N., and S. Ueno, "Requirements of an MPLS Transport Profile", RFC 5654, September 2009.

9.2. Informative References

- [I-D.ietf-ccamp-assoc-info]
Berger, L., "Usage of The RSVP Association Object", draft-ietf-ccamp-assoc-info-02 (work in progress), May 2011.
- [I-D.ietf-ccamp-asymm-bw-bidir-lsps-bis]
Takacs, A., Berger, L., Caviglia, D., Fedyk, D., and J. Meuric, "GMPLS Asymmetric Bandwidth Bidirectional Label Switched Paths (LSPs)", draft-ietf-ccamp-asymm-bw-bidir-lsps-bis-01 (work in progress), January 2011.
- [I-D.ietf-ccamp-mpls-tp-cp-framework]
Andersson, L., Berger, L., Fang, L., Bitar, N., Gray, E., Takacs, A., Vigoureux, M., and E. Bellagamba, "MPLS-TP Control Plane Framework", draft-ietf-ccamp-mpls-tp-cp-framework-06 (work in progress), February 2011.

- [RFC2205] Braden, B., Zhang, L., Berson, S., Herzog, S., and S. Jamin, "Resource ReSerVation Protocol (RSVP) -- Version 1 Functional Specification", RFC 2205, September 1997.
- [RFC3209] Awduche, D., Berger, L., Gan, D., Li, T., Srinivasan, V., and G. Swallow, "RSVP-TE: Extensions to RSVP for LSP Tunnels", RFC 3209, December 2001.
- [RFC3473] Berger, L., "Generalized Multi-Protocol Label Switching (GMPLS) Signaling Resource ReserVation Protocol-Traffic Engineering (RSVP-TE) Extensions", RFC 3473, January 2003.
- [RFC5511] Farrel, A., "Routing Backus-Naur Form (RBNF): A Syntax Used to Form Encoding Rules in Various Routing Protocol Specifications", RFC 5511, April 2009.
- [RFC5920] Fang, L., "Security Framework for MPLS and GMPLS Networks", RFC 5920, July 2010.

Authors' Addresses

Fei Zhang (editor)
ZTE

Email: zhang.feiz3@zte.com.cn

Ruiquan Jing
China Telecom

Email: jingrq@ctbri.com.cn

Fan Yang
ZTE

Email: yang.fan5@zte.com.cn

Weilian Jiang
ZTE

Email: jiang.weilian@zte.com.cn

MPLS Working Group
Internet-Draft
Intended status: Standards Track
Expires: January 12, 2012

E. Bellagamba, Ed.
L. Andersson
Ericsson
P. Skoldstrom, Ed.
Acreo AB
D. Ward
J. Drake
Juniper
July 11, 2011

Configuration of Pro-Active Operations, Administration, and Maintenance
(OAM) Functions for MPLS-based Transport Networks using LSP Ping
draft-ietf-mpls-lsp-ping-mpls-tp-oam-conf-02

Abstract

This specification describes the configuration of pro-active MPLS-TP Operations, Administration, and Maintenance (OAM) Functions for a given LSP using a set of TLVs that are carried by the LSP Ping protocol

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

Status of this Memo

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

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

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

This Internet-Draft will expire on January 12, 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	3
1.1.	Requirements Language	4
2.	Overview of MPLS OAM for Transport Applications	4
3.	Theory of Operations	5
3.1.	MPLS OAM Configuration Operation Overview	5
3.1.1.	Configuration of BFD sessions	5
3.1.2.	Configuration of Performance Monitoring	6
3.1.3.	Configuration of Measurements and FMS	6
3.2.	OAM Functions TLV	6
3.2.1.	BFD Configuration sub-TLV	8
3.2.1.1.	Local Discriminator sub-TLV	10
3.2.1.2.	Negotiation Timer Parameters sub-TLV	10
3.2.1.3.	BFD Authentication sub-TLV	11
3.2.2.	MPLS OAM Source MEP-ID sub-TLV	12
3.2.3.	Performance Monitoring sub-TLV	13
3.2.3.1.	MPLS OAM PM Loss sub-TLV	14
3.2.3.2.	MPLS OAM PM Delay sub-TLV	15
3.2.4.	MPLS OAM FMS sub-TLV	16
3.3.	IANA Considerations	17
4.	OAM configuration errors	17
5.	Security Considerations	18
6.	References	18
6.1.	Normative References	18
6.2.	Informative References	19
	Authors' Addresses	20

1. Introduction

This document describes the configuration of pro-active MPLS-TP Operations, Administration, and Maintenance (OAM) Functions for a given LSP using TLVs carried in LSP Ping [BFD-Ping]. In particular it specifies the mechanisms necessary to establish MPLS-TP OAM entities for monitoring and performing measurements on an LSP, as well as defining information elements and procedures to configure pro-active MPLS OAM functions. Initialization and control of on-demand MPLS 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 three options for configuring MPLS-TP OAM, without a control plane by either using an NMS or LSP Ping, or with a control plane using GMPLS (specifically RSVP-TE) .

Pro-active MPLS OAM is performed by three different protocols, Bidirectional Forwarding Detection (BFD) [RFC5880] for Continuity Check/Connectivity Verification, the delay measurement protocol (DM) [MPLS-PM] for delay and delay variation (jitter) measurements, and the loss measurement protocol (LM) [MPLS-PM] for packet loss and throughput measurements. Additionally there is a number of Fault Management Signals that can be configured.

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 track the liveliness and detect data plane failures of MPLS-TP point-to-point and might also be extended to support point-to-multipoint connections.

The delay and loss measurements protocols [MPLS-PM] use a simple query/response model for performing bidirectional measurements that allows the originating node to measure packet loss and delay in both directions. By timestamping and/or writing current packet counters to the measurement packets at four times (Tx and Rx in both directions) current delays and packet losses can be calculated. By performing successive delay measurements the 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.

MPLS Transport Profile (MPLS-TP) describes a profile of MPLS that enables operational models typical in transport networks, while providing additional OAM, survivability and other maintenance functions not currently supported by MPLS. [RFC5860] defines the requirements for the OAM functionality 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 PWE3 architectures to support the capabilities and functionalities of a packet transport network.

1.1. Requirements Language

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

2. Overview of MPLS OAM for Transport Applications

[MPLS-TP-OAM-FWK] describes how MPLS OAM mechanisms are operated to meet transport requirements outlined in [RFC5860].

[BFD-CCCV] specifies two BFD operation modes: 1) "CC mode", which uses periodic BFD message exchanges with symmetric timer settings, supporting Continuity Check, 2) "CV/CC mode" which sends unique maintenance entity identifiers in the periodic BFD messages supporting Connectivity Verification as well as Continuity Check.

[MPLS-PM] specifies mechanisms for performance monitoring of LSPs, in particular it specifies loss and delay measurement OAM functions.

[MPLS-FMS] specifies fault management signals with which a server LSP can notify client LSPs about various fault conditions to suppress alarms or to be used as triggers for actions in the client LSPs. The following signals are defined: Alarm Indication Signal (AIS), Link Down Indication (LDI) and Locked Report (LKR). To indicate client faults associated with the attachment circuits Client Signal Failure Indication (CSF) can be used. CSF is described in [MPLS-TP-OAM-FWK] and in the context of this document is for further study.

[MPLS-TP-OAM-FWK] describes the mapping of fault conditions to consequent actions. Some of these mappings may be configured by the operator, depending on the application of the LSP. The following defects are identified: Loss Of Continuity (LOC), Misconnectivity, MEP Misconfiguration and Period Misconfiguration. Out of these defect conditions, the following consequent actions may be

configurable: 1) whether or not the LOC defect should result in blocking the outgoing data traffic; 2) whether or not the "Period Misconfiguration defect" should result in a signal fail condition.

3. Theory of Operations

3.1. MPLS OAM Configuration Operation Overview

LSP Ping, or alternatively RSVP-TE [RSVP-TE CONF], can be used to simply enable the different OAM functions, by setting the corresponding flags in the "OAM Functions TLV". Additionally one may include sub-TLVs for the different OAM functions in order to specify different parameters in detail.

3.1.1. Configuration of BFD sessions

For this specification, BFD MUST be run in either one of the two modes:

- Asynchronous mode, where both sides should be in active mode.
- Unidirectional mode

In the simplest scenario LSP Ping, or alternatively RSVP-TE [RSVP-TE CONF], 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 G-ACh 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" 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. actual authentication not only error detection) the "BFD Authentication sub-TLV" 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.

3.1.2. Configuration of Performance Monitoring

It is possible to configure Performance Monitoring functionalities such as Loss, Delay and Throughput as described in [MPLS-PM].

When configuring Performance monitoring functionalities it can be chosen either the default configuration (by only setting the respective flags in the "OAM functions TLV") or a customized configuration (by including the respective Loss and/or Delay sub-TLVs).

3.1.3. Configuration of Measurements and FMS

Additional OAM functions may be configured by setting the appropriate flags in the "OAM Functions TLV", these include Performance Measurements (packet loss, throughput, delay, and delay variation) and Fault Management Signal handling.

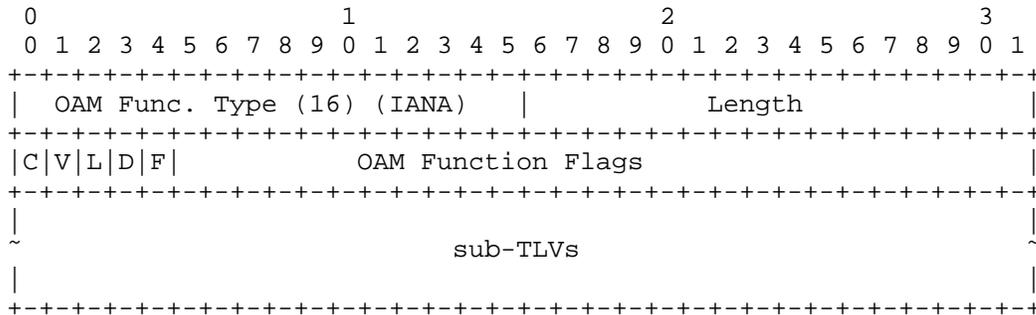
By setting the PM Loss flag in the "OAM Functions TLV" and including the "MPLS OAM PM Loss sub-TLV" one can configure the measurement interval and loss threshold values for triggering protection.

Delay measurements are configured by setting PM Delay flag in the "OAM Functions TLV" and including the "MPLS OAM PM Loss sub-TLV" one can configure the measurement interval and the delay threshold values for triggering protection.

To configure Fault Monitoring Signals and their refresh time the FMS flag in the "OAM Functions TLV" MUST be set and the "MPLS OAM FMS sub-TLV" included.

3.2. OAM Functions TLV

The "OAM Functions TLV" depicted below is carried as a TLV of the LSP Echo request/response messages.



The "OAM Functions TLV" contains a number of flags indicating which OAM functions should be activated as well as OAM function specific sub-TLVs with configuration parameters for the particular function.

Type: indicates a new type, the "OAM Functions TLV" (IANA to define, suggested value 16).

Length: the length of the OAM Function Flags field including the total length of the sub-TLVs in octets.

OAM Function Flags: a bitmap numbered from left to right as shown in the figure.

These flags are defined in this document:

OAM Function Flag bit#	Description
0 (C)	Continuity Check (CC)
1 (V)	Connectivity Verification (CV)
2 (F)	Fault Management Signals (FMS)
3 (L)	Performance Monitoring/Loss (PM/Loss)
4 (D)	Performance Monitoring/Delay (PM/Delay)
5 (T)	Throughput Measurement
6-31	Reserved (set all to 0s)

Sub-TLVs corresponding to the different flags are as follows:

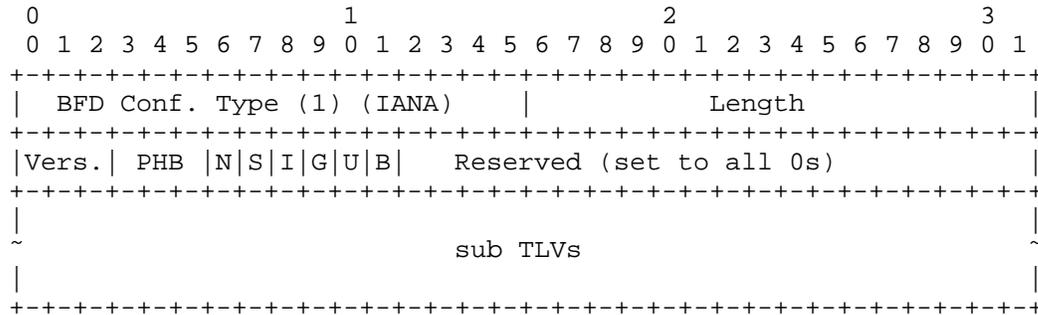
- "BFD Configuration sub-TLV", which MUST be included if the CC and/or the CV OAM Function flag is set. This sub-TLV MUST carry a "BFD Local Discriminator sub-TLV" and a "Timer Negotiation Parameters sub-TLV" if the N flag is cleared. "MPLS OAM Source MEP-ID sub-TLV" MUST also be included. If the I flag is set, the "BFD Authentication sub-TLV" may be included.

- "MPLS OAM PM Loss sub-TLV" within the "Performance Monitoring sub-TLV", which MAY be included if the PM/Loss OAM Function flag is set. If the "MPLS OAM 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 measurement interval different from the default ones. In fact the throughput measurement make use of the same tool as the loss measurement, hence the same TLV is used.
- "MPLS OAM PM Delay sub-TLV" within the "Performance Monitoring sub-TLV", which MAY be included if the PM/Delay OAM Function flag is set. If the "MPLS OAM PM Delay sub-TLV" is not included, default configuration values are used.
- "MPLS OAM FMS sub-TLV", which MAY be included if the FMS OAM Function flag is set. If the "MPLS OAM FMS sub-TLV" is not included, default configuration values are used.

3.2.1. BFD Configuration sub-TLV

The "BFD Configuration sub-TLV" (depicted below) 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.



Type: indicates a new type, the "BFD Configuration sub-TLV" (IANA to define, suggested value 1).

Length: indicates the length of the TLV including sub-TLVs but excluding the Type and Length 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".

PHB: Identifies the Per-Hop Behavior (PHB) to be used for periodic continuity monitoring messages.

BFD Negotiation (N): If set timer negotiation/re-negotiation via BFD Control Messages is enabled, when cleared it is disabled.

Symmetric session (S): If set the BFD session MUST use symmetric timing values.

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).

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 UDP packets. If both the G bit and U bit are set, configuration gives precedence to the G bit.

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.

The "BFD Configuration sub-TLV" MUST include the following sub-TLVs in the LSP Echo request message:

- "Local Discriminator sub-TLV";
- "Negotiation Timer Parameters sub-TLV" if the N flag is cleared.

The "BFD Configuration sub-TLV" MUST include the following sub-TLVs in the LSP Echo reply message:

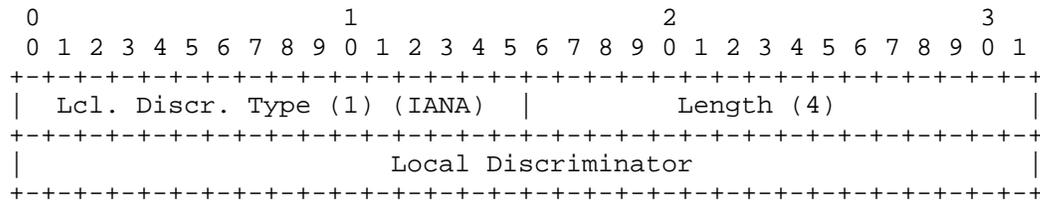
- "Local Discriminator sub-TLV";
- "Negotiation Timer Parameters sub-TLV" if:
 - the N and S flags are cleared
 - the N flag is cleared and the S flag is set and a timing interval larger than the one received needs to be used

Reserved: Reserved for future specification and set to 0.

3.2.1.1. Local Discriminator sub-TLV

The "Local Discriminator sub-TLV" is carried as a sub-TLV of the "BFD Configuration sub-TLV" and is depicted below.

[Author's note: This should be aligned with RFC5884, exactly how to do that is under discussion.]



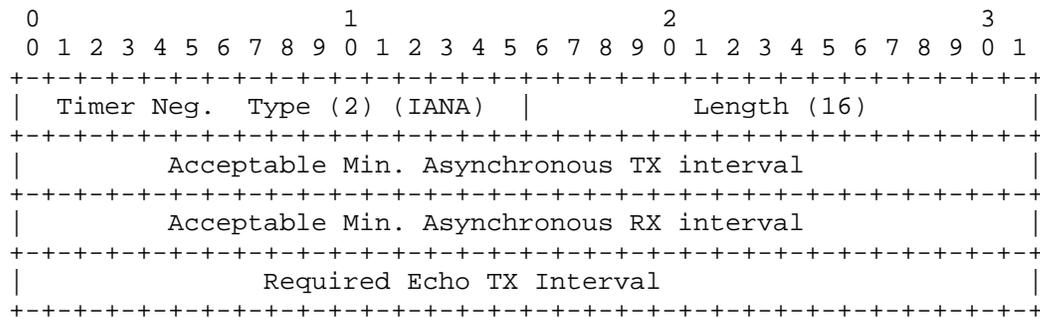
Type: indicates a new type, the "Local Discriminator sub-TLV" (IANA to define, suggested value 1).

Length: indicates the TLV total length in octets.

Local Discriminator: A unique, nonzero discriminator value generated by the transmitting system and referring to itself, used to demultiplex multiple BFD sessions between the same pair of systems.

3.2.1.2. 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 below.



Type: indicates a new type, the "Negotiation Timer Parameters sub-TLV" (IANA to define, suggested value 2).

Length: indicates the length of the parameters in octets (16).

Acceptable Min. Asynchronous TX interval: in case of S (symmetric) flag set in the "BFD Configuration" TLV, it expresses the desired time interval (in microseconds) at which the LER initiating the signaling intends to both transmit and receive BFD periodic control packets. If the receiving edge LSR can not support such value, it is allowed to reply back 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.

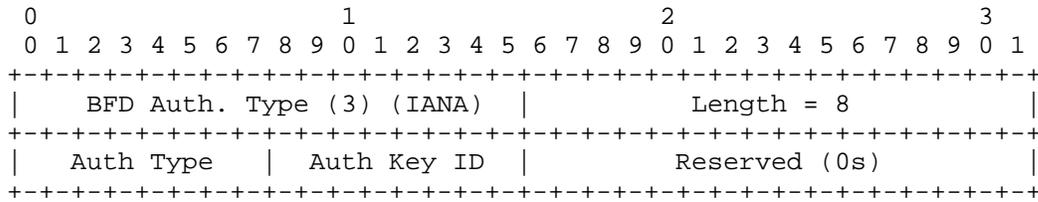
Acceptable Min. Asynchronous RX interval: in case of S (symmetric) flag set in the "BFD Configuration sub-TLV", this field MUST be equal to "Acceptable Min. Asynchronous TX interval" and has no additional meaning respect to the one described for "Acceptable Min. Asynchronous TX interval".

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 "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 can not support this value an error MUST be generated "Unsupported BFD TX Echo rate interval". By default the value is set to 0.

3.2.1.3. BFD Authentication sub-TLV

The "BFD Authentication sub-TLV" is carried as a sub-TLV of the "BFD Configuration sub-TLV" and is depicted below.



Type: indicates a new type, the "BFD Authentication sub-TLV" (IANA to define).

Length: indicates the TLV total length in octets. (8)

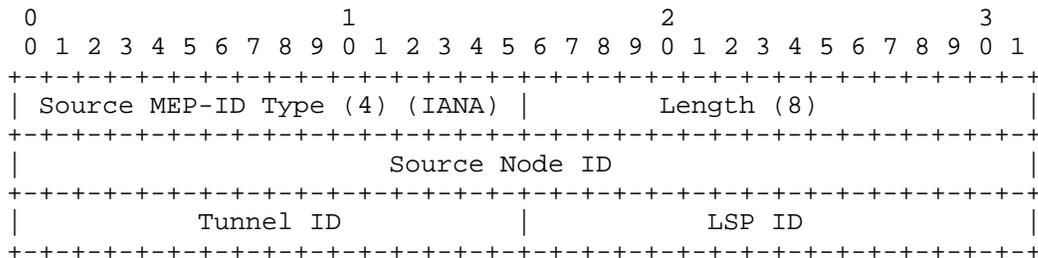
Auth Type: indicates which type of authentication to use. The same values as are defined in section 4.1 of [RFC5880] are used.

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.

Reserved: Reserved for future specification and set to 0.

3.2.2. MPLS OAM Source MEP-ID sub-TLV

The "MPLS OAM Source MEP-ID sub-TLV" depicted below is carried as a sub-TLV of the "OAM Functions TLV".



Type: indicates a new type, the "MPLS OAM Source MEP-ID sub-TLV" (IANA to define, suggested value 3).

Length: indicates the length of the parameters in octets (8).

Source Node ID: 32-bit node identifier as defined in [MPLS-TP-IDENTIF].

Tunnel ID: a 16-bit unsigned integer unique to the node as defined in [MPLS-TP-IDENTIF].

LSP ID: a 16-bit unsigned integer unique within the Tunnel_ID as defined in [MPLS-TP-IDENTIF].

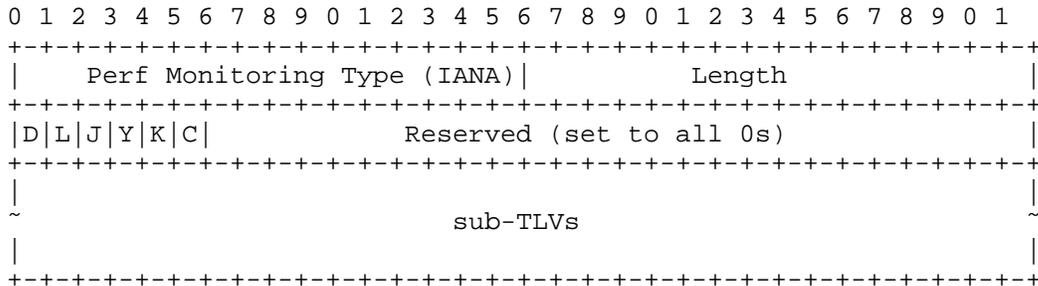
3.2.3. Performance Monitoring sub-TLV

If the "OAM functions TLV" has either the L (Loss), D (Delay) or T (Throughput) flag set, the "Performance Monitoring sub-TLV" MUST be present.

In case the vlues needs to be different than the default ones the "Performance Monitoring sub-TLV", "MPLS OAM PM Loss sub-TLV" MAY include the following sub-TLVs:

- "MPLS OAM PM Loss sub-TLV" if the L flag is set in the "OAM functions TLV";
- "MPLS OAM PM Delay sub-TLV" if the D flag is set in the "OAM functions TLV";

The "Performance Monitoring sub-TLV" depicted below is carried as a sub-TLV of the "OAM Functions TLV".



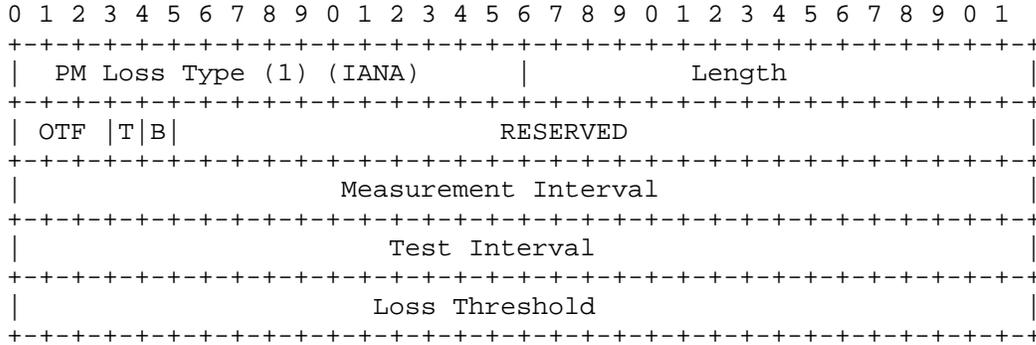
Configuration Flags, for the specific function description please refer to [MPLS-PM]:

- D: Delay inferred/direct (0=INFERRED, 1=DIRECT)
- L: Loss inferred/direct (0=INFERRED, 1=DIRECT)
- J: Delay variation/jitter (1=ACTIVE, 0=NOT ACTIVE)
- Y: Dyadic (1=ACTIVE, 0=NOT ACTIVE)
- K: Loopback (1=ACTIVE, 0=NOT ACTIVE)

- C: Combined (1=ACTIVE, 0=NOT ACTIVE)

3.2.3.1. MPLS OAM PM Loss sub-TLV

The "MPLS OAM PM Loss sub-TLV" depicted below is carried as a sub-TLV of the "Performance Monitoring sub-TLV".



Type: indicates a new type, the "MPLS OAM PM Loss sub-TLV" (IANA to define, suggested value 1).

Length: indicates the length of the parameters in octets (12).

OTF: Origin Timestamp Format of the Origin Timestamp field described in [MPLS-PM]. By default it is set to IEEE 1588 version 1.

Configuration Flags, please refer to [MPLS-PM] 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.

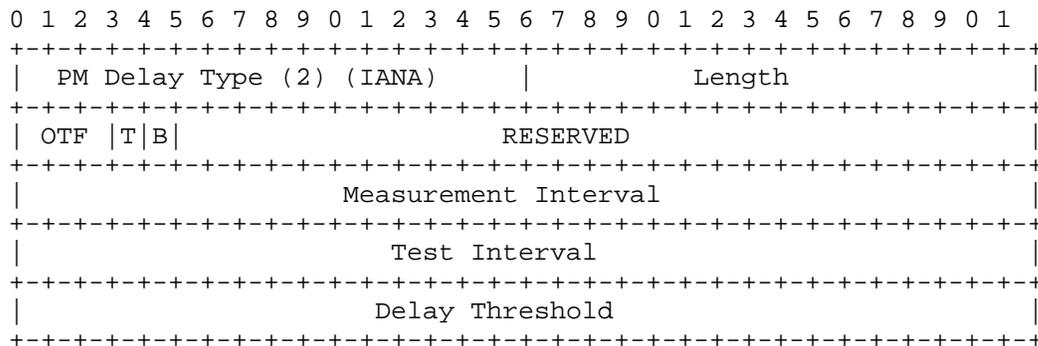
Measurement Interval: the time interval (in microseconds) at which Loss Measurement query messages MUST be sent on both directions. If the edge LSR receiving the Path message can not support such value, it can reply back with a higher interval. By default it is set to (TBD).

Test Interval: test messages interval as described in [MPLS-PM]. By default it is set to (TBD).

Loss Threshold: the threshold value of lost packets over which protections MUST be triggered. By default it is set to (TBD).

3.2.3.2. MPLS OAM PM Delay sub-TLV

The "MPLS OAM PM Delay sub-TLV" depicted below is carried as a sub-TLV of the "OAM Functions TLV".



Type: indicates a new type, the "MPLS OAM PM Loss sub-TLV" (IANA to define, suggested value 1).

Length: indicates the length of the parameters in octets (12).

OTF: Origin Timestamp Format of the Origin Timestamp field described in [MPLS-PM]. By default it is set to IEEE 1588 version 1.

Configuration Flags, please refer to [MPLS-PM] 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.

Measurement Interval: the time interval (in microseconds) at which

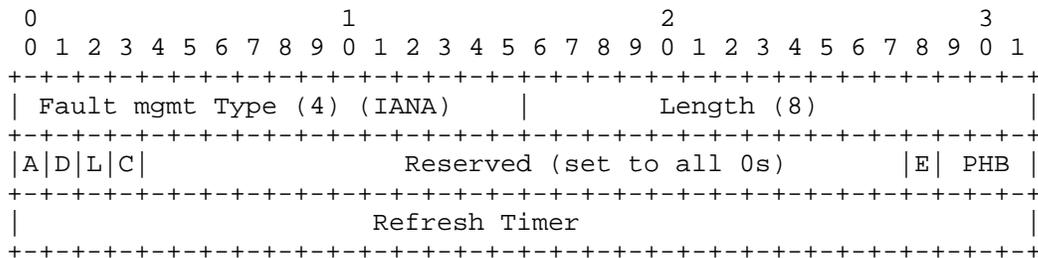
Delay Measurement query messages MUST be sent on both directions. If the edge LSR receiving the Path message can not support such value, it can reply back with a higher interval. By default it is set to (TBD).

Test Interval: test messages interval as described in [MPLS-PM]. By default it is set to (TBD).

Delay Threshold: the threshold value of measured delay (in microseconds) over which protections MUST be triggered. By default it is set to (TBD).

3.2.4. MPLS OAM FMS sub-TLV

The "MPLS OAM FMS sub-TLV" depicted below is carried as a sub-TLV of the "OAM Functions TLV".



Type: indicates a new type, the "MPLS OAM FMS sub-TLV" (IANA to define, suggested value 4).

Length: indicates the length of the parameters in octets (8).

Signal Flags: are used to enable the following signals:

- A: Alarm Indication Signal (AIS) as described in [MPLS-FMS]
- D: Link Down Indication (LDI) as described in [MPLS-FMS]
- L: Locked Report (LKR) as described in [MPLS-FMS]
- C: Client Signal Failure (CSF) as described in [MPLS-CSF]
- Remaining bits: Reserved for future specification and set to 0.

Configuration Flags:

- E: used to enable/disable explicitly clearing faults
- PHB: identifies the per-hop behavior of packets with fault management information

Refresh Timer: indicates the refresh timer (in microseconds) of fault indication messages. If the edge LSR receiving the Path message can not support such value, it can reply back with a higher interval.

3.3. IANA Considerations

This document specifies the following new TLV types:

- "OAM Functions" type: 16;

sub-TLV types to be carried in the "OAM Functions TLV":

- "BFD Configuration" type: 1;
- "MPLS OAM PM Loss" type: 2;
- "MPLS OAM PM Delay" type: 3;
- "MPLS OAM FMS" type: 4.

sub-TLV types to be carried in the "BFD Configuration sub-TLV":

- "Local Discriminator" type: 1;
- "Negotiation Timer Parameters" type: 2;
- "MPLS OAM Source MEP-ID" type: 3.
- "BFD Authentication" sub-TLV type: 4

4. OAM configuration errors

This document specifies additional Return Codes to LSP Ping:

- "MPLS OAM Unsupported Functionality" (IANA to assign, suggested value 16);
- "OAM Problem/Unsupported TX rate interval" (IANA to assign, suggested value 17).

5. Security Considerations

The signaling of OAM related parameters and the automatic establishment of OAM entities introduces additional security considerations to those discussed in [RFC3473]. 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.

Security aspects will be covered in more detailed in subsequent versions of this document.

6. References

6.1. Normative References

[MPLS-FMS]

Swallow, G., Fulignoli, A., Vigoureux, M., Boutros, S., and D. Ward, "MPLS Fault Management OAM", 2009, <draft-ietf-mpls-tp-fault>.

[MPLS-PM]

Bryant, S. and D. Frost, "Packet Loss and Delay Measurement for the MPLS Transport Profile", 2010, <draft-ietf-mpls-loss-delay>.

[MPLS-PM-Profile]

Bryant, S. and D. Frost, "A Packet Loss and Delay Measurement Profile for MPLS-based Transport Networks", 2010, <draft-ietf-mpls-tp-loss-delay-profile>.

[MPLS-TP-IDENTIF]

Bocci, M., Swallow, G., and E. Gray, "MPLS-TP Identifiers", 2010, <draft-ietf-mpls-tp-identifiers>.

[OAM-CONF-FWK]

Takacs, A., Fedyk, D., and J. van He, "OAM Configuration Framework for GMPLS RSVP-TE", 2009, <draft-ietf-ccamp-oam-configuration-fwk>.

[RFC2119]

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

[RFC3471]

Berger, L., "Generalized Multi-Protocol Label Switching (GMPLS) Signaling Functional Description", RFC 3471, January 2003.

[RFC5586]

Bocci, M., Vigoureux, M., and S. Bryant, "MPLS Generic

Associated Channel", RFC 5586, June 2009.

- [RFC5654] Niven-Jenkins, B., Brungard, D., Betts, M., Sprecher, N., and S. Ueno, "Requirements of an MPLS Transport Profile", RFC 5654, September 2009.
- [RFC5860] Vigoureux, M., Ward, D., and M. Betts, "Requirements for Operations, Administration, and Maintenance (OAM) in MPLS Transport Networks", RFC 5860, May 2010.
- [RFC5880] Katz, D. and D. Ward, "Bidirectional Forwarding Detection (BFD)", RFC 5880, June 2010.
- [RSVP-TE CONF] Bellagamba, E., Ward, D., Andersson, L., and P. Skoldstrom, "Configuration of pro-active MPLS-TP Operations, Administration, and Maintenance (OAM) Functions Using RSVP-TE", 2010, <draft-ietf-ccamp-rsvp-te-mpls-tp-oam-ext>.

6.2. Informative References

- [BFD-CCCV] Allan, D., Swallow, G., and J. Drake, "Proactive Connectivity Verification, Continuity Check and Remote Defect indication for MPLS Transport Profile", 2010, <draft-ietf-mpls-tp-bfd-cc-cv-rdi-03>.
- [BFD-Ping] Bahadur, N., Aggarwal, R., Ward, D., Nadeau, T., Sprecher, N., and Y. Weingarten, "LSP Ping and BFD encapsulation over ACH", 2010, <draft-ietf-mpls-tp-lsp-ping-bfd-procedures-01>.
- [ETH-OAM] Takacs, A., Gero, B., Fedyk, D., Mohan, D., and D. Long, "GMPLS RSVP-TE Extensions for Ethernet OAM", 2009, <draft-ietf-ccamp-rsvp-te-eth-oam-ext>.
- [MPLS-TP OAM Analysis] Sprecher, N., Weingarten, Y., and E. Bellagamba, "MPLS-TP OAM Analysis", 2011, <draft-ietf-mpls-tp-oam-analysis>.
- [MPLS-TP-OAM-FWK] Bocci, M. and D. Allan, "Operations, Administration and Maintenance Framework for MPLS-based Transport Networks", 2010, <draft-ietf-mpls-tp-oam-framework>.
- [RFC3479] Farrel, A., "Fault Tolerance for the Label Distribution

Protocol (LDP)", RFC 3479, February 2003.

[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.

[RFC5921] Bocci, M., Bryant, S., Frost, D., Levrau, L., and L. Berger, "A Framework for MPLS in Transport Networks", RFC 5921, July 2010.

Authors' Addresses

Elisa Bellagamba (editor)
Ericsson
Torshamnsgatan 48
Kista, 164 40
Sweden

Email: elisa.bellagamba@ericsson.com

Loa Andersson
Ericsson
Torshamnsgatan 48
Kista, 164 40
Sweden

Phone:
Email: loa.andersson@ericsson.com

Pontus Skoldstrom (editor)
Acreo AB
Electrum 236
Kista, 164 40
Sweden

Phone: +46 8 6327731
Email: pontus.skoldstrom@acreo.se

Dave Ward
Juniper

Phone:
Email: dward@juniper.net

John Drake
Juniper

Phone:
Email: jdrake@juniper.net

Network Working Group
Internet Draft
Intended status: Standards Track
Expires: December 5, 2011

Sami Boutros (Ed.)
Siva Sivabalan (Ed.)
Cisco Systems, Inc.

Rahul Aggarwal (Ed.)
Juniper Networks, Inc.

Martin Vigoureux (Ed.)
Alcatel-Lucent

Xuehui Dai (Ed.)
ZTE Corporation

June 5, 2011

MPLS Transport Profile Lock Instruct and Loopback Functions
draft-ietf-mpls-tp-li-lb-02.txt

Status of this Memo

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

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

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

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

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

This Internet-Draft will expire on December 5, 2011.

Abstract

This document specifies an extension to MPLS Operation, administration, and Maintenance (OAM) to operate an Label Switched Path (LSP), bi-directional RSVP-TE tunnels, Pseudowires (PW), or Multi-segment PWS in loopback mode for management purpose in an MPLS based Transport. This extension includes mechanism to lock and unlock MPLS-TP Tunnels (i.e. data and control traffic) and can be used to loop all traffic (i.e, data and control traffic) at a specified LSR on the path of the LSP in an MPLS based Transport Network back to the source. However, the mechanisms are intended to be applicable to other aspects of MPLS as well.

Table of Contents

1. Introduction.....3
2. Terminology.....5
3. Loopback/Lock Mechanism.....5
3.1. In-band Message Identification.....5
3.2. LI-LB Message Format.....6
3.3. Return codes.....7
3.4. Cause codes.....7
3.5. Authentication TLV.....8
3.6. LSP Ping Extensions.....9
3.6.1. LI-LB Request TLV.....9
3.6.2. LI-LB Response TLV.....9
4. Loopback/Lock Operations.....9
4.1. Lock Request.....10
4.2. Unlock Request.....10
4.3. Loopback Request.....10
4.4. Loopback Removal.....11
5. Data packets.....11
6. Operation.....11
6.1. General Procedures.....11
6.2. Example Topology.....11
6.3. Locking an LSP.....12
6.4. Unlocking an LSP.....13
6.5. Setting an LSP into Loopback mode.....14
6.6. Removing an LSP from Loopback mode.....15
7. Security Considerations.....16
8. IANA Considerations.....16
8.1. Pseudowire Associated Channel Type.....16
8.2. New LSP Ping TLV types.....16
9. Acknowledgements.....16
10. References.....16
10.1. Normative References.....16
10.2. Informative References.....17
Author's Addresses.....17
Full Copyright Statement.....19
Intellectual Property Statement.....19

In traditional transport networks, circuits are provisioned across multiple nodes and service providers have the ability to operate the transport circuit such as T1 line in loopback mode for management purposes, e.g., to test or verify connectivity of the circuit up to a specific node on the path of the circuit, to test the circuit performance with respect to delay/jitter, etc. This document provides the same loopback capability for the bi-directional LSPs in MPLS based Transport Networks emulating traditional transport circuits. The mechanisms in this document apply to co-routed bidirectional paths as defined in [7], which include LSPs, bi-directional RSVP-TE tunnels, Pseudowires (PW), and Multi-segment PWs in MPLS based Transport Networks. However, the mechanisms are intended to be applicable to other aspects of MPLS as well.

This document specifies how to operate the Lock and Loopback functions over both the Generic Associated Channel (GACH) and over LSP-Ping. LSP-Ping itself can run either over the GACH or using native IP addressing; the manner in which LSP-Ping is transported in an MPLS-TP network is out of the scope of this document.

This document uses a sample topology to describe the lock instruct and loopback functions. This sample topology comprises four MPLS-TP nodes [A---B---C---D]. There is an LSP from A to D, and thus A and D are MEPs and B and C are MIPs. Unless otherwise specified, the operator desires to lock the LSP (this is done on A and D, by definition) and loop the LSP at C.

That is, the desired behavior is that all packets transmitted by A on this locked and looped LSP arrive at C from B and are encapsulated in the D->A direction by C such that these packets reach A.

Locking and looping an LSP is a two-step process. The first step is to lock the LSP so that it is not made available to carry user traffic. The locking of an LSP is managed by the two MEPs of an LSP - in this example, A and D. Locking is controlled by one of the MEPs; this example uses A. A sends a Lock request message to D along the LSP, either in the GACH or in LSP-Ping. This message will be received by D as it is the far-end MEP for that LSP. D responds to the lock request with an ACK or NACK; the ACK indicates that D has taken the LSP out of service (i.e. Locked the LSP) and the NACK indicates that D cannot comply with the Lock request. In general, if a message (e.g. Lock request, Loopback request) cannot be complied with, the node which received the request replies with a NACK and a cause code; the details of error message processing are discussed later in this document.

Once A has received the ACK to its Lock request, A is then allowed to put the LSP in Loopback mode. In order to set the LSP in Loopback mode, A sends a Loopback request message to the MIP or MEP where A desired the loopback to be enabled. In this example, A desires to set the loopback at C, although note that it is possible to A to set the loopback at any node downstream of A (e.g. B, C, D). The TTL on the Loopback request message is set by A such that the TTL expires when it reaches the node where A wants the loopback to be set (in this case, C). C responds to the Loopback request with a reply message (ACK/NACK) back to A to indicate whether it has successfully set the LSP into the Loopback mode.

If A receives an ACK from its Loopback request, the LSP is now in Loopback mode. A is free to send any test packets down this LSP as it sees fit. These packets MUST NOT be forwarded towards D. As the LSP is locked, D MUST NOT transmit any traffic on the LSP in the reverse direction (that is, D->A). Any traffic received by C from the reverse direction MUST be dropped and MAY be logged, as the receipt of traffic by C in the D->A direction indicates an error.

When A desires to remove the LSP from Loopback state, it begins to reverse the Loopback and Lock. This is a two-step process; first A removes the Loopback from C, then A removes the Lock from D. This process is similar to the process of establishing Lock and Loopback in the first place. A sends a Loopback Remove message to C using the TTL method described above, and C ACKs or NACKs the Loopback Remove. Once A receives the Loopback Remove ACK from C, A sends a Lock Remove message to D. D must ACK or NACK this message. Once A receives the Lock Remove ACK from D, the LSP is brought back into normal operation.

The proposed mechanism is based on a new set of messages and TLVs which can be transported using one of the following methods:

- (1) An in-band MPLS message transported using a new ACH code point, the message will have different types to perform the loopback request/remove and Lock/unlock functions, and may carry new set of TLVs.
- (2) A new set of TLVs which can be transported using LSP-Ping extensions defined in [4], and in compliance to specifications [5].

Method (1) and (2) are referred to as "in-band option" and "LSP-Ping option" respectively in the rest of the document.

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].

2. Terminology

- ACH: Associated Channel Header
- LSR: Label Switching Router
- MEP: Maintenance Entity Group End Point
- MIP: Maintenance Entity Group Intermediate Point.
- MPLS-TP: MPLS Transport Profile
- MPLS-OAM: MPLS Operations, Administration and Maintenance
- MPLS-TP LSP: Bidirectional Label Switch Path representing a circuit
- NMS: Network Management System
- TLV: Type Length Value
- TTL: Time To Live
- LI-LB: Lock instruct-Loopback

3. Loopback/Lock Mechanism

For the in-band option, the proposed mechanism uses a new code point in the Associated Channel Header (ACH) described in [6].

3.1. In-band Message Identification

In the in-band option, the LI-LB channel is identified by the ACH as defined in RFC 5586 [6] with the Channel Type set to the LI-LB code point = 0xHH. [HH to be assigned by IANA from the PW Associated Channel Type registry] The LI-LB Channel does not use ACH TLVs and MUST not include the ACH TLV header. The LI-LB ACH Channel is shown below.

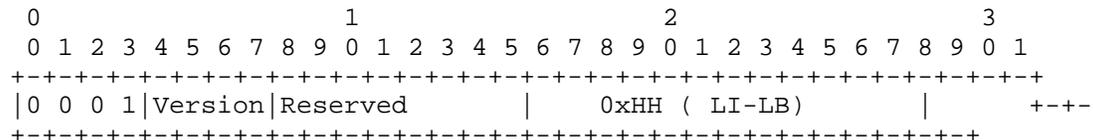


Figure 1: ACH Indication of LI-LB

The LI-LB Channel is 0xHH (to be assigned by IANA)

The format of an LI-LB Message is shown below.

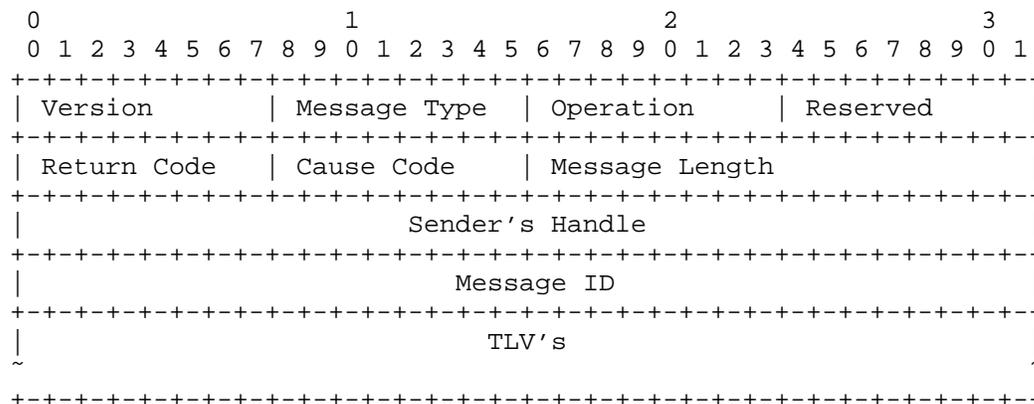


Figure 2: MPLS LI-LB Message Format

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

Message Type

Two message types are defined as shown below.

Message Type	Description
0x0	LI-LB request
0x1	LI-LB response

Operation

Four operations are defined as shown below. The operations can appear in a Request or Response message.

Operation	Description
0x1	Lock
0x2	Unlock
0x3	Set_Loopback

Message Length

The total length of any included TLVs.

Sender's Handle

The Sender's Handle is filled in by the sender, and MUST be copied unchanged by the receiver in the MPLS response message (if any). There are no semantics associated with this handle, although a sender may find this useful for matching up requests with replies.

Message ID

The Message ID is set by the sender of an MPLS request message. It MUST be copied unchanged by the receiver in the MPLS response message (if any). A sender SHOULD increment this value on each new message. A retransmitted message SHOULD leave the value unchanged.

The Return code and Cause code only have meaning in a Response message. In a request message the Return code and Cause code must be set to zero and ignored on receipt. Return codes and cause codes are described in the following Sections.

3.3. Return codes

Value	Meaning
-----	-----
0	Informational
1	Success
2	Failure

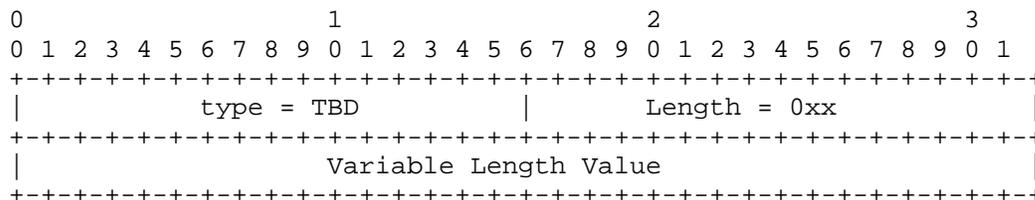
3.4. Cause codes

Value	Meaning
-----	-----
0	Success
1	Fail to match target MIP/MEP ID
2	Malformed LI-LB request received
3	One or more of the TLVs is/are unknown
4	Authentication failed
5	LSP/PW already locked
6	LSP/PW already unlocked
7	Fail to lock LSP/PW

- 8 Fail to unlock LSP/PW
- 9 LSP/PW already in loopback mode
- 10 LSP/PW is not in loopback mode
- 11 Fail to set LSP/PW in loopback mode
- 12 Fail to remove LSP/PW from loopback mode
- 13 No label binding for received message
- 14 Authentication required but not received.

Note that in the case of cause code 3, the unknown TLV can also be optionally included in the response. For failure responses with multiple causes only the first cause code can be included.

3.5. Authentication TLV



The PPP CHAP described in [9] will be used to authenticate the LI-LB request.

The variable length value carried in the optional authentication TLV, will include the Packet Format described in section 3.2 of [9].

The optional authentication TLV can be included in the MPLS OAM LSP Ping echo messages containing a LI-LB request TLV or in the inband LI-LB Message. When an authentication TLV is present in the Request message the CHAP procedures described in section 3.2 of [9] MUST be followed.

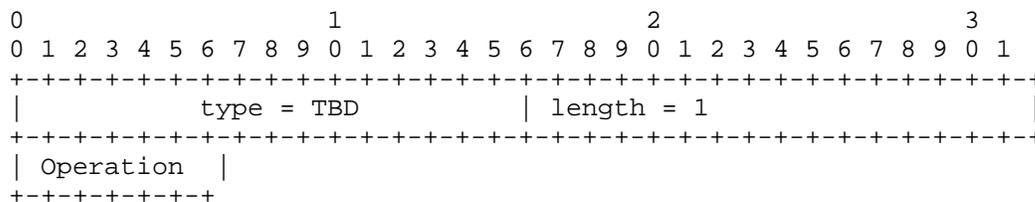
The CHAP packets will be transmitted by the authenticator using LI-LB Request or response messages, responses to the authentication protocol messages will be transmitted using LI-LB request or response messages.

If the CHAP negotiation results in a failure, the authenticator or the sender of the request message MUST stop requesting the LI-LB function.

A receiver of an LI-LB request, MAY send an error "Authentication required but not received", if the optional authentication TLV is not included in the LI-LB request.

3.6. LSP Ping Extensions

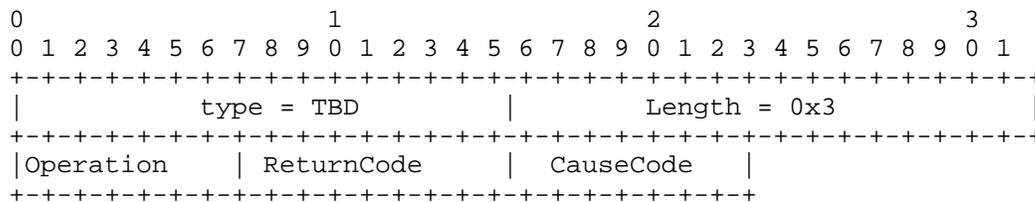
3.6.1. LI-LB Request TLV



Operation	Description
0x1	Lock
0x2	Unlock
0x3	Set_Loopback
0x4	Unset_Loopback

A MEP includes a LI-LB Request TLV in the MPLS LSP Ping echo request message to request the MEP on the other side of the LSP to perform Lock/Unlock and Set/Unset Loopback operations. Only one LI-LB request TLV can be present in an LSP Ping Echo request message.

3.6.2. LI-LB Response TLV



Only one LI-LB response TLV can be present in an LSP Ping Echo request message.

4. Loopback/Lock Operations

When performing a Lock or Loopback function, the reply to a message MUST use the same method as the original message. That is, if a node requests lock or loopback using LSP Ping then any replies to that request must also use LSP Ping; if a node requests lock or loopback using in-band, any replies to that request must use in-band. It is permissible to use different methods for the lock and the loopback functions on a given LSP. For example, a node can lock an LSP using the LSP Ping method and then can loop the LSP using the in-band method, or vice versa.

An ACK response of a request will be a response message with return code 1 (success) and cause code 0, while a NACK response will have a return code 2 (failure) and the corresponding cause code.

4.1. Lock Request

Lock Request is used to request a MEP to take an LSP out of service so that some form of maintenance can be done.

The receiver MEP MUST send either an ACK or a NAK response to the sender MEP. Until the sender MEP receives an ACK, it MUST NOT assume that the receiver MEP has taken the LSP out of service. A receiver MEP sends an ACK only if it can successfully lock the LSP. Otherwise, it sends a NAK.

The receiver MEP once locked, MUST discard all received traffic.

4.2. Unlock Request

The Unlock Request is sent from the MEP which has previously sent lock request. Upon receiving the unlock request message, the receiver MEP brings the LSP back in service.

The receiver MEP MUST send either an ACK or a NAK response to the sender MEP. Until the sender MEP receives an ACK, it MUST NOT assume that the LSP has been put back in service. A receiver MEP sends an ACK only if the LSP has been unlocked, and unlock operation is successful. Otherwise, it sends a NAK.

4.3. Loopback Request

When a MEP wants to put an LSP in loopback mode, it sends a Loopback request message. The message can be intercepted by either a MIP or a MEP depending on the MPLS TTL value. The receiver puts in corresponding LSP in loopback mode.

The receiver MEP or MIP MUST send either an ACK or NAK response to the sender MEP. An ACK response is sent if the LSP is successfully put in loopback mode. Otherwise, a NAK response is sent. Until an ACK response is received, the sender MEP MUST NOT assume that the LSP can operate in loopback mode.

When loopback mode operation of an LSP is no longer required, the MEP that previously sent the Loopback request message sends another Loopback Removal message. The receiver MEP changes the LSP from loopback mode to normal mode of operation.

The receiver MEP or MIP MUST send either an ACK or NAK response to the sender MEP. An ACK response is sent if the LSP is already in loopback mode, and if the LSP is successfully put back in normal operation mode. Otherwise, a NAK response is sent. Until an ACK response is received, the sender MEP MUST NOT assume that the LSP is put back in normal operation mode.

5. Data packets

Data packets sent from the sender MEP will be looped back to that sender MEP. OAM packets not intercepted by TTL expiry will as well be looped back. The use of data packets to measure packet loss, delay and delay variation is outside the scope of this document.

6. Operation

6.1. General Procedures

When placing an LSP into Loopback mode, the operation MUST first be preceded by a Lock operation.

When sending Loopback Request/Removal using LSP Ping or in-Band messages the TTL of the topmost label is set as follows:-

If the target node is a MIP, the TTL MUST be set to the exact number of hops required to reach that MIP.

If the target node is a MEP, the value MUST be set to at least the number of hops required to reach that MEP. For most operations where the target is a MEP, the TTL MAY be set to 255.

However, to remove a MEP from Loopback mode, the sending MEP MUST set the TTL to the exact number of hops required to reach the MEP (if the TTL were set higher, the Loopback removal message would be looped back toward the sender).

6.2. Example Topology

The next four sections discuss the procedures for Locking, Unlocking, setting an LSP into loopback, and removing the loopback. The description is worded using an example. Assume an LSP traverses nodes A <--> B <--> C <--> D. We will refer to the Maintenance Entities

Internet-Draft draft-ietf-mpls-tp-li-lb-02.txt June 2011
involved as MEP-A, MIP-B, MIP-C, and MEP-D respectively. Suppose a maintenance operation invoked at MEP-A requires a loopback be set at MIP-C. To invoke Loopack mode at MIP-C, A would first need to lock the LSP. Then it may proceed to set the loopback at C. Following the loopback operation, A would need to remove the loopback at C and finally unlock the LSP.

The following sections describe MEP-A setting and unsetting a lock at MEP-D and then setting and removing a loopback at MIP-C.

6.3. Locking an LSP

1. MEP-A sends an MPLS LSP Ping Echo request message with the Lock TLV or an in-Band Lock request Message. Optionally, an authentication TLV MAY be included.

2. Upon receiving the request message, D uses the received label stack and the Target Stack FEC TLV as per [5]/source MEP-ID to identify the LSP. If no label binding exists or there is no associated LSP back to the originator, the event is logged. Processing ceases. Otherwise the message is delivered to the target MEP.

- a. if the source MEP-ID does not match, the event is logged and processing ceases.

- b. if the target MEP-ID does not match, MEP-D sends a failure response with cause code 1.

MEP-D then examines the message, and:

- c. if the message is malformed, it sends a failure response with cause code 2 back to MEP-A.

- d. if message authentication fails, it MAY send a failure response with cause code 4 back to MEP-A.

- e. if any of the TLVs is not known, it sends a failure response with cause code 3 back to MEP-A. It may also include the unknown TLVs.

- f. if the LSP is already locked, it sends a response with cause code 5 back to MEP-A.

- g. if the LSP is not already locked and cannot be locked, it sends a failure response with cause code 7 back to A.

- h. if the LSP is successfully locked, it sends a success response with cause code 0 (Success) back to MEP-A.

The response is sent using an MPLS LSP Ping echo reply with a response TLV or an in-Band Lock response message. An authentication TLV MAY be included.

MEP-D will lock the LSP, resulting in that all traffic from D to A, including all OAM traffic, stops.

- a. MEP-A will detect a discontinuation in the OAM traffic, e.g. cv and cc packets, but since it has been informed that the LSP will be locked it will take no action(s).
- b. When MEP-A receives the LI ACK, MEP-A discontinues sending other OAM traffic, e.g. cv and cc packets. MEP-D will detect this, but since it is in Locked state it will take no action.

6.4. Unlocking an LSP

1. MEP-A sends an MPLS Echo request message with the unLock TLV or an in-Band unLock request Message. Optionally, an authentication TLV MAY be included.

2. Upon receiving the unLock request message, D uses the received label stack and target FEC/source MEP-ID as per [5] to identify the LSP. If no label binding exists or there is no associated LSP back to the originator, the event is logged. Processing ceases. Otherwise the message is delivered to the target MEP.

a. if the source MEP-ID does not match, the event is logged and processing ceases.

b. if the target MEP-ID does not match, MEP-D sends a failure response with cause code 1.

MEP-D then examines the message, and:

c. if the message is malformed, it sends a failure response with cause code 2 back to MEP-A.

d. if message authentication fails, it MAY send a failure response with cause code 4 back to MEP-A.

e. if any of the TLVs is not known, it sends a failure response with cause code 3 back to MEP-A. It may also include the unknown TLVs.

f. if the LSP is already unlocked, it sends a response with cause code 6 back to MEP-A.

g. if the LSP is locked and cannot be unlocked, it sends a response with cause code 8 back to MEP-A.

h. if the LSP is successfully unlocked, it sends a success response with cause code 0 (Success) back to MEP-A.

The response is sent using an MPLS LSP Ping echo reply with a response TLV or an in-Band unlock response message. An authentication TLV MAY be included.

6.5. Setting an LSP into Loopback mode

1. MEP-A sends an MPLS LSP Ping Echo request message with the loopback TLV or an in-Band Loopback request message. Optionally, an authentication TLV MAY be included.

2. Upon intercepting the MPLS Loopback message via TTL expiration, C uses the received label stack and target FEC/source MEP-ID as per [5] to identify the LSP.

If no label binding exists or there is no associated LSP back to the originator, the event is logged. Processing ceases.

Otherwise the message is delivered to the target MIP/MEP - in this case MIP-C.

a. if the source MEP-ID does not match, the event is logged and processing ceases.

b. if the target MIP-ID does not match, MIP-C sends a failure response with cause code 1.

MIP-C then examines the message, and:

c. if the message is malformed, it sends a failure response with cause code 2 back to MEP-A.

d. if the message authentication fails, it sends a failure response with cause code 4 back to MEP-A.

e. if any of the TLV is not known, C sends a failure response with cause code 3 back to MEP-A. It may also include the unknown TLVs.

f. if the LSP is already in the requested loopback mode, it sends a failure response with cause code 9 back to MEP-A.

g. if the LSP is not already in the requested loopback mode and that loopback mode cannot be set, it sends a failure response with cause code 11 back to MEP-A.

h. if the LSP is successfully programmed into the requested loopback mode, it sends a success response with cause code 0 (Success) back to MEP-A.

The response is sent using an MPLS LSP Ping echo reply with a response TLV or an in-Band Loopback response message. An authentication TLV MAY be included.

6.6. Removing an LSP from Loopback mode

1. MEP-A sends a MPLS LSP Ping Echo request message with the Loopback removal TLV or an in-Band Loopback removal request message. Optionally, an authentication TLV MAY be included.

2. Upon intercepting the MPLS Loopback removal message via TTL expiration, C uses the received label stack and the target FEC/source MEP-ID as per [5] to identify the LSP.

If no label binding exists or there is no associated LSP back to the originator, the event is logged. Processing ceases.

Otherwise the message is delivered to the target MIP/MEP - in this case MIP-C.

a. if the source MEP-ID does not match, the event is logged and processing ceases.

b. if the target MIP-ID does not match, MIP-C sends a failure response with cause code 1 back to MEP-A.

MIP-C then examines the message, and:

c. if the message is malformed, it sends a failure response with cause code 2 back to MEP-A.

d. if the message authentication fails, it sends a failure response with cause code 4 back to MEP-A.

e. if any of the TLV is not known, C sends a failure response with cause code 3 back to MEP-A. It may also include the unknown TLVs.

f. if the LSP is not in loopback mode, it sends a failure response with cause code 10 back to MEP-A.

g. if the LSP loopback cannot be removed, it sends a failure response with cause code 12 back to MEP-A.

h. if the LSP is successfully changed from loopback mode to normal mode of operation, it sends a reply with cause code 0 (Success) back to MEP-A.

The response is sent using an MPLS LSP Ping echo reply with a response TLV or an in-Band Loopback removal response message. An authentication TLV MAY be included.

7. Security Considerations

Security is addressed through the use of authentication TLV and the the Challenge-Handshake Authentication protocol procedures described in section [9].

8. IANA Considerations

8.1. Pseudowire Associated Channel Type

LI-LB OAM requires a unique Associated Channel Type which is assigned by IANA from the Pseudowire Associated Channel Types Registry.

Registry:

Value	Description	TLV Follows	Reference
-----	-----	-----	-----
0xHHHH	LI-LB	No	(Section 3.1)

8.2. New LSP Ping TLV types

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

1. LI-LB Request TLV (See section 3.3.1)
2. LI-LB Response TLV (See section 3.3.2)
3. Authentication TLV (See section 3.3.3)

9. Acknowledgements

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

10. References

10.1. Normative References

- [1] Niven-Jenkins, B., Brungard, D., Betts, M., Sprecher, N., and S. Ueno, "Requirements of an MPLS Transport Profile", RFC 5654, September 2009.
- [2] Vigoureux, M., Ward, D., and M. Betts, "Requirements for Operations, Administration, and Maintenance (OAM) in MPLS Transport Networks", RFC 5860, May 2010.
- [3] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

- [4] K. Kompella, G. Swallow, "Detecting Multi-Protocol Label Switched (MPLS) Data Plane Failures", RFC 4379, February 2006.
- [5] N. Bahadur, et. al., "MPLS on-demand Connectivity Verification, Route Tracing and Adjacency Verification", draft-ietf-mpls-tp-on-demand-cv-00, work in progress, June 2010
- [6] Bocci, M., Vigoureux, M., and S. Bryant, "MPLS Generic Associated Channel", RFC 5586, June 2009.
- [7] Bocci, M. and G. Swallow, "MPLS-TP Identifiers", draft-ietf-mpls-tp-identifiers-01 (work in progress), June 2010.
- [8] Niven-Jenkins, B., Brungard, D., Betts, M., Sprecher, N., and S.Ueno, "Requirements of an MPLS Transport Profile", RFC 5654, September 2009.
- [9] B. Lloyd, L&A, and W. Simpson, "PPP Authentication Protocols", October 1992.

10.2. Informative References

- [10] Nabil Bitar, et. al, "Requirements for Multi-Segment Pseudowire Emulation Edge-to-Edge (PWE3) ", RFC5254, October 2008.

Author's Addresses

Sami Boutros
Cisco Systems, Inc.
Email: sboutros@cisco.com

Siva Sivabalan
Cisco Systems, Inc.
Email: msiva@cisco.com

Rahul Aggarwal
Juniper Networks.
EMail: rahul@juniper.net

Martin Vigoureux
Alcatel-Lucent.
Email: martin.vigoureux@alcatel-lucent.com

Xuehui Dai
ZTE Corporation.
Email: dai.xuehui@zte.com.cn

George Swallow
Cisco Systems, Inc.
Email: swallow@cisco.com

David Ward
Juniper Networks.
Email: dward@juniper.net

Stewart Bryant
Cisco Systems, Inc.
Email: stbryant@cisco.com

Carlos Pignataro
Cisco Systems, Inc.
Email: cpignata@cisco.com

Eric Osborne
Cisco Systems, Inc.
Email: eosborne@cisco.com

Nabil Bitar
Verizon.
Email: nabil.bitar@verizon.com

Italo Busi
Alcatel-Lucent.
Email: italo.busi@alcatel-lucent.it

Lieven Levrau
Alcatel-Lucent.
Email: llevrau@alcatel-lucent.com

Laurent Ciavaglia
Alcatel-Lucent.
Email: laurent.ciavaglia@alcatel-lucent.com

Bo Wu
ZTE Corporation.
Email: wu.bo@zte.com.cn

Full Copyright Statement

Copyright (c) 2008 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.

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.

Intellectual Property Statement

The IETF 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 this 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.

Acknowledgment

Funding for the RFC Editor function is currently provided by the Internet Society.

Internet Engineering Task Force
Internet-Draft
Intended status: Informational
Expires: November 17, 2011

L. Fang, Ed.
Cisco Systems Inc.
B. Niven-Jenkins, Ed.
Velocix
S. Mansfield, Ed.
Ericsson
May 16, 2011

MPLS-TP Security Framework
draft-ietf-mpls-tp-security-framework-01

Abstract

This document provides a security framework for Multiprotocol Label Switching Transport Profile (MPLS-TP). Extended from MPLS technologies, MPLS-TP introduces new OAM capabilities, transport oriented path protection mechanism, and strong emphasis on static provisioning supported by network management systems. This document addresses the security aspects that are relevant in the context of MPLS-TP specifically. It describes the security requirements for MPLS-TP; potential securities threats and migration procedures for MPLS-TP networks and MPLS-TP inter-connection to MPLS and GMPLS networks.

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

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.]

Status of this Memo

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

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

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

This Internet-Draft will expire on November 17, 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	4
1.1.	Background and Motivation	4
1.2.	Scope	4
1.3.	Requirement Language	5
1.4.	Terminology	6
1.5.	Structure of the document	7
2.	Security Reference Models	7
2.1.	Security Reference Model 1	7
2.2.	Security Reference Model 2	9
2.3.	Security Reference Model 3	12
2.4.	Trusted Zone Boundaries	13
3.	Security Requirements for MPLS-TP	14
4.	Security Threats	16
4.1.	Attacks on the Control Plane	18
4.2.	Attacks on the Data Plane	18
5.	Defensive Techniques for MPLS-TP Networks	19
5.1.	Authentication	19
5.1.1.	Management System Authentication	19
5.1.2.	Peer-to-Peer Authentication	20
5.1.3.	Cryptographic Techniques for Authenticating Identity	20
5.2.	Access Control Techniques	20
5.3.	Use of Isolated Infrastructure	21
5.4.	Use of Aggregated Infrastructure	21
5.5.	Service Provider Quality Control Processes	21
5.6.	Verification of Connectivity	21
6.	Monitoring, Detection, and Reporting of Security Attacks	21
7.	Security Considerations	22
8.	IANA Considerations	22
9.	References	22
9.1.	Normative References	22
9.2.	Informative References	23
	Authors' Addresses	23

1. Introduction

1.1. Background and Motivation

This document provides a security framework for Multiprotocol Label Switching Transport Profile (MPLS-TP).

MPLS-TP Requirements and MPLS-TP Framework are defined in [RFC5654] and [RFC5921] respectively. The intent of MPLS-TP development is to address the needs for transport evolution, the fast growing bandwidth demand accelerated by new packet based services and multimedia applications, from Ethernet Services, Layer 2 and Layer 3 VPNS, triple play to Mobile Access Network (RAN) backhaul, etc. MPLS-TP is based on MPLS technologies to take advantage of the technology maturity, and it is required to maintain the transport characteristics.

Focused on meeting transport requirements, MPLS-TP uses a subset of MPLS features, and introduces extensions to reflect the transport technology characteristics. The added functionalities include in-band OAM, transport oriented path protection and recovery mechanisms, etc. There is strong emphasis on static provisioning supported by Network Management System (NMS) or Operation Support System (OSS). There are also needs for MPLS-TP and MPLS interworking.

The security aspects for the new extensions which are particularly designed for MPLS-TP need to be addressed. The security models, requirements, threat and defense techniques previously defined in [RFC5921] can be used for the re-use of the existing functionalities in MPLS and GMPLS, but not sufficient to cover the new extensions.

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

1.2. Scope

This document addresses the security aspects that are specific to MPLS-TP. It intends to provide the security requirements for MPLS-TP; define security models which apply to various MPLS-TP deployment scenarios; identify the potential security threats and mitigation procedures for MPLS-TP networks and MPLS-TP inter-connection to MPLS or GMPLS networks. Inter-AS and Inter-provider security for MPLS-TP to MPLS-TP connections or MPLS-TP to MPLS connections are discussed, where connections present higher security risk factors than connections for Intra-AS MPLS-TP.

The general security analysis and guidelines for MPLS and GMPLS are addressed in [RFC5920], the content which has no new impact to MPLS-TP will not be repeated in this document. Other general security issues regarding transport networks that are not specific to MPLS-TP are also out of scope. Readers may also refer to the "Security Best Practices Efforts and Documents" Opsec Effort [opsec-efforts] and "Security Mechanisms for the Internet" [RFC3631] (if there are linkages to the Internet in the applications) for general network operation security considerations. This document does not intend to define the specific mechanisms/methods that must be implemented to satisfy the security requirements.

Issues/Areas to be addressed:

- o G-Ach (control plane attack, DoS attack, message intercept, etc.)
- o Spoofing ID
- o Loopback
- o NMS attack
- o NMS and CP interaction
- o MIP/MEP assignment and attacks
- o Topology discovery
- o Data plane authentication
- o Label authentication
- o DoS attack in Data Plane
- o Performance Monitoring

1.3. Requirement Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in [RFC2119]. Although this document is not a protocol specification, the use of this language clarifies the instructions to protocol designers producing solutions that satisfy the requirements set out in this document.

1.4. Terminology

This document uses MPLS, MPLS-TP, and Security specific terminology. Detailed definitions and additional terminology for MPLS-TP may be found in [RFC5654], [RFC5921], and MPLS/GMPLS security related terminology in [RFC5920].

- o BFD: Bidirectional Forwarding Detection
- o CE: Customer-Edge device
- o DoS: Denial of Service
- o DDoS: Distributed Denial of Service
- o GAL: Generic Alert Label
- o G-ACH: Generic Associated Channel
- o GMPLS: Generalized Multi-Protocol Label Switching
- o LDP: Label Distribution Protocol
- o LSP: Label Switched Path
- o MCC: Management Communication Channel
- o MEP: Maintenance End Point
- o MIP: Maintenance Intermediate Point
- o MPLS: MultiProtocol Label Switching
- o OAM: Operations, Administration, and Management
- o PE: Provider-Edge device
- o PSN: Packet-Switched Network
- o PW: Pseudowire
- o RSVP: Resource Reservation Protocol
- o RSVP-TE: Resource Reservation Protocol with Traffic Engineering Extensions
- o S-PE: Switching Provider Edge

- o SSH: Secure Shell
- o TE: Traffic Engineering
- o TLS: Transport Layer Security
- o T-PE: Terminating Provider Edge
- o VPN: Virtual Private Network
- o WG: Working Group of IETF
- o WSS: Web Services Security

1.5. Structure of the document

Section 1: Introduction

Section 2: MPLS-TP Security Reference Models

Section 3: Security Requirements

Section 4: Security Threats

Section 5: Defensive/Mitigation techniques/procedures

2. Security Reference Models

This section defines a reference model for security in MPLS-TP networks.

The models are built on the architecture of MPLS-TP defined in [RFC5921]. The Service Provider (SP) boundaries play an important role in determining the security models for any particular deployment.

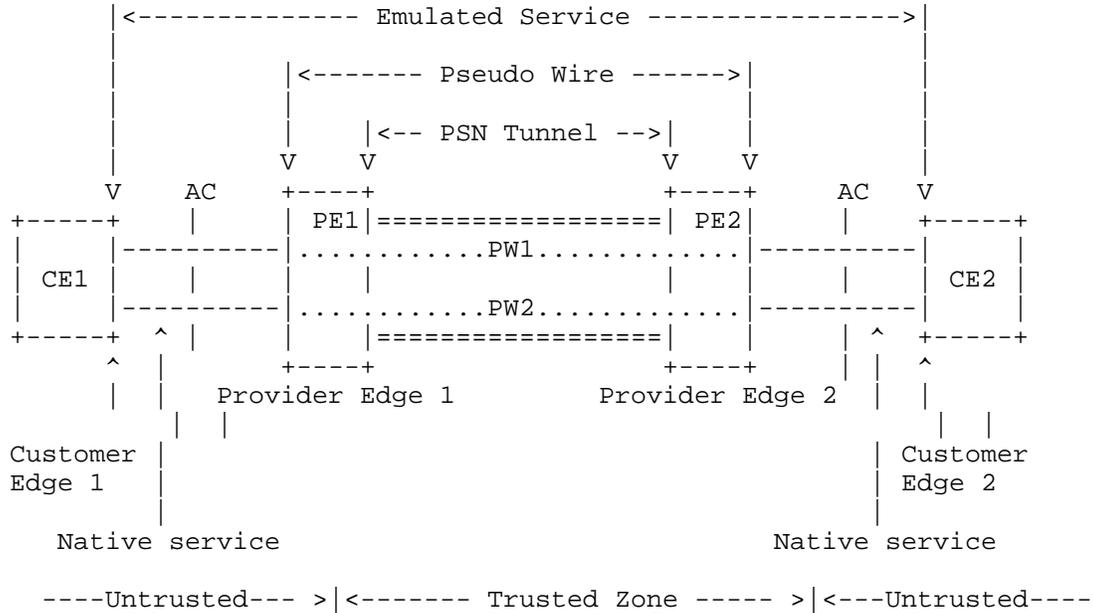
This document defines a trusted zone as being where a single SP has the total operational control over that part of the network. A primary concern is about security aspects that relate to breaches of security from the "outside" of a trusted zone to the "inside" of this zone.

2.1. Security Reference Model 1

In the reference model 1, a single SP has total control of PE/T-PE to PE/T-PE of the MPLS-TP network.

Security reference model 1(a)

An MPLS-TP network with Single Segment Pseudowire (SS-PW) from PE to PE. The trusted zone is PE1 to PE2 as illustrated in MPLS-TP Security Model 1 (a) (Figure 1).

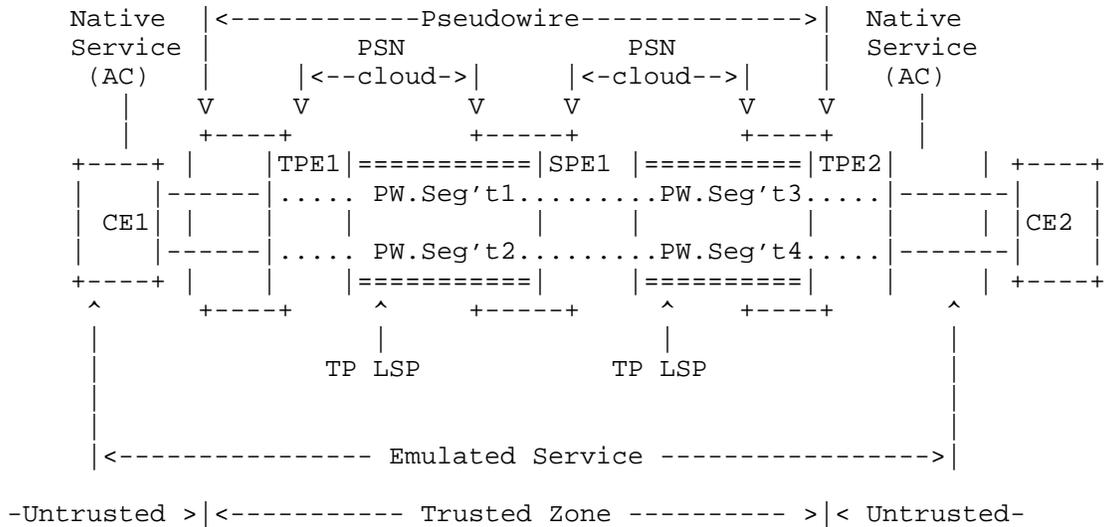


MPLS-TP Security Model 1 (a)

Figure 1

Security reference model 1(b)

An MPLS-TP network with Multi-Segment Pseudowire (MS-PW) from T-PE to T-PE. The trusted zone is T-PE1 to T-PE2 in this model as illustrated in MPLS-TP Security Model 1 (b) (Figure 2).



MPLS-TP Security Model 1 (b)

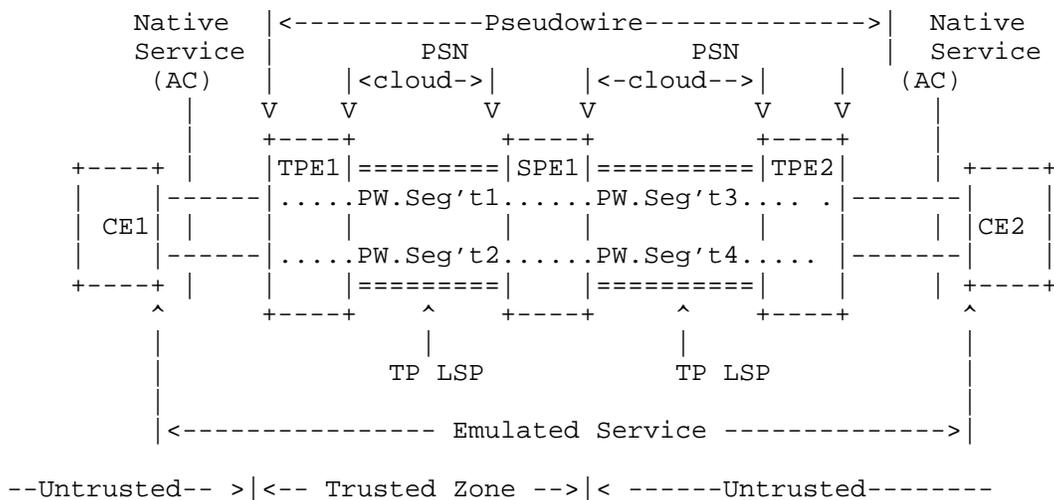
Figure 2

2.2. Security Reference Model 2

In the reference model 2, a single SP does not have the total control of PE/T-PE to PE/T-PE of the MPLS-TP network, S-PE and T-PE may be under the control of different SPs or their customers or may not be trusted for some other reason. The MPLS-TP network is not contained within a single trusted zone.

Security Reference Model 2(a)

An MPLS-TP network with Multi-Segment Pseudowire (MS-PW) from T-PE to T-PE. The trusted zone is T-PE1 to S-PE, as illustrated in MPLS-TP Security Model 2 (a) (Figure 3).

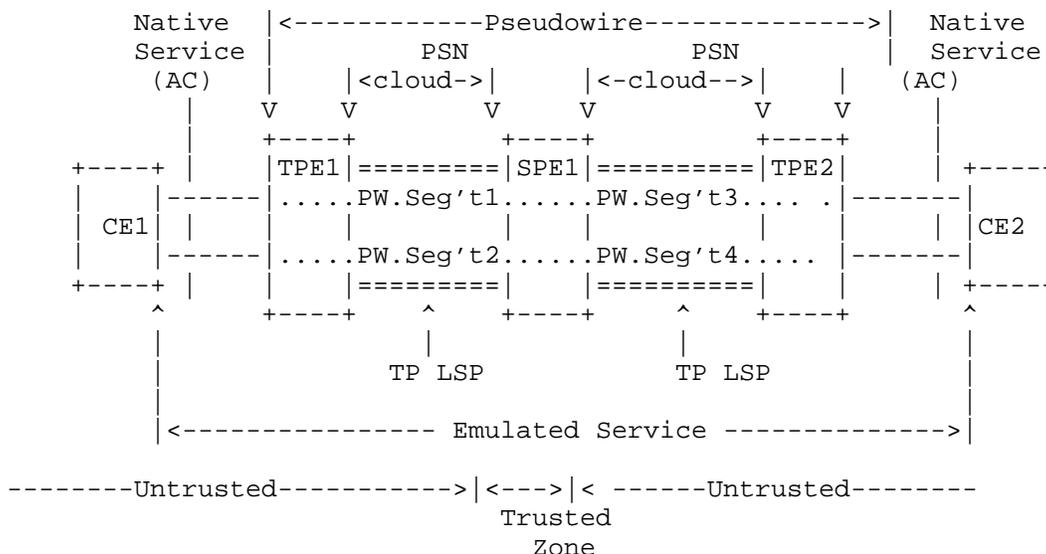


MPLS-TP Security Model 2 (a)

Figure 3

Security Reference Model 2(b)

An MPLS-TP network with Multi-Segment Pseudowire (MS-PW) from T-PE to T-PE. The trusted zone is the S-PE, as illustrated in MPLS-TP Security Model 2 (b) (Figure 4).

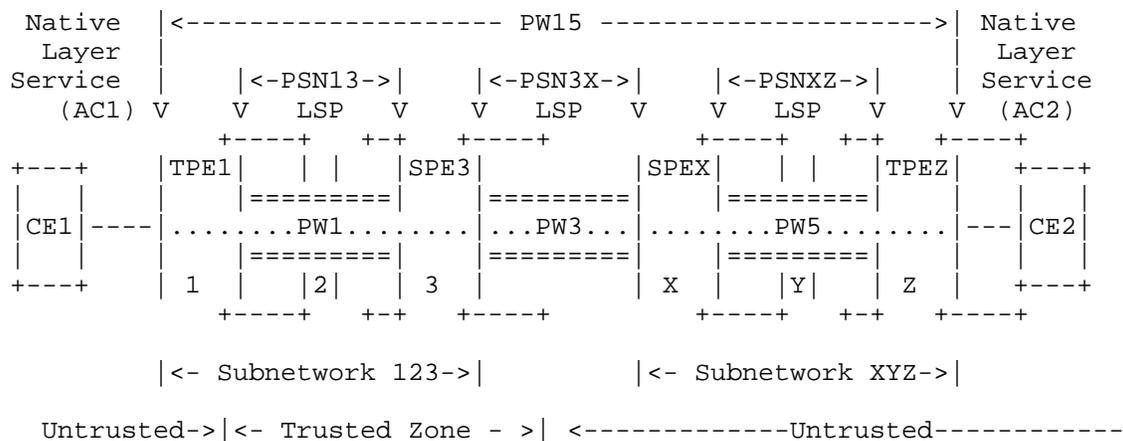


MPLS-TP Security Model 2 (b)

Figure 4

Security Reference Model 2(c)

An MPLS-TP network with Multi-Segment Pseudowire (MS-PW) from different Service Providers with inter-provider PW connections. The trusted zone is T-PE1 to S-PE3, as illustrated in MPLS-TP Security Model 2 (c) (Figure 5).

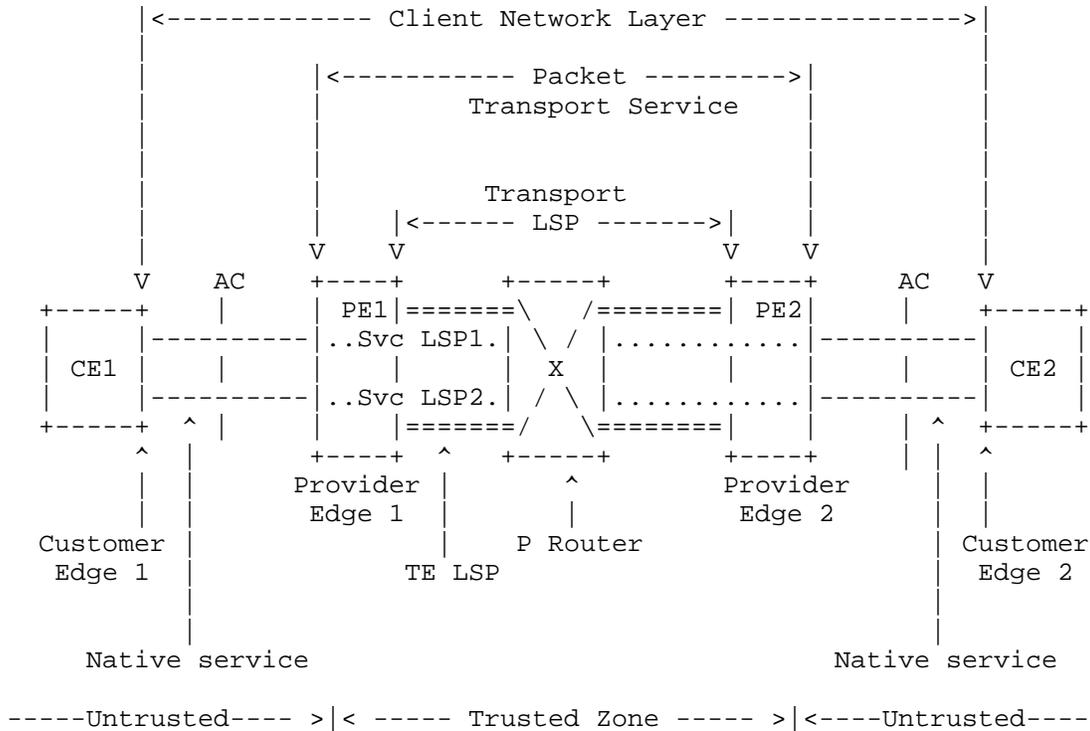


MPLS-TP Security Model 2 (c)

Figure 5

2.3. Security Reference Model 3

An MPLS-TP network with a Transport LSP from PE1 to PE2. The trusted zone is PE1 to PE2 as illustrated in MPLS-TP Security Model 3 (a) (Figure 6).



MPLS-TP Security Model 3 (a)

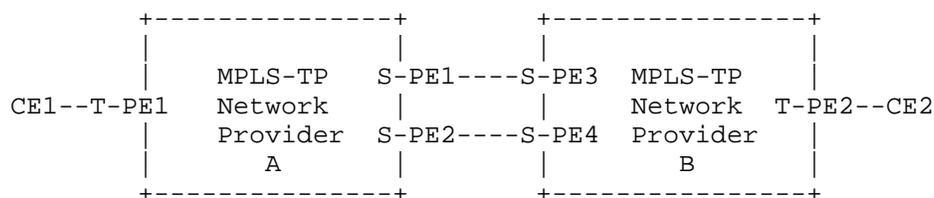
Figure 6

2.4. Trusted Zone Boundaries

The boundaries of a trusted zone should be carefully defined when analyzing the security properties of each individual network, as illustrated from the above, the security boundaries determine which reference model should be applied to the use case analysis.

A key requirement of MPLS-TP networks is that the security of the trusted zone MUST NOT be compromised by interconnecting one SP's MPLS-TP or MPLS infrastructure with another SP's core, T-PE devices, or end users.

In addition, neighboring nodes in the network may be trusted or untrusted. Neighbors may also be authorized or unauthorized. Even though a neighbor may be authorized for communication, it may not be trusted. For example, when connecting with another provider's S-PE to set up Inter-AS LSPs, the other provider is considered to be untrusted but may be authorized for communication.



For Provider A:

Trusted Zone: Provider A MPLS-TP network
 Trusted neighbors: T-PE1, S-PE1, S-PE2
 Authorized but untrusted neighbor: Provider B
 Unauthorized neighbors: CE2

MPLS-TP trusted zone and authorized neighbor

Figure 7

3. Security Requirements for MPLS-TP

This section covers security requirements for securing MPLS-TP network infrastructure. The MPLS-TP network can be operated without a control plane or via dynamic control planes protocols. The security requirements related to new MPLS-TP OAM, recovery mechanisms, MPLS-TP and MPLS interconnection, and MPLS-TP specific operational requirements will be addressed in this section.

A service provider may choose the implementation options which are the best fit for his/her network operation. This document does not state that a MPLS/GMPLS network must fulfill all security requirements listed to be secure.

These requirements are focused on: 1) how to protect the MPLS-TP network from various attacks originating outside the trusted zone including those from network users, both accidental and malicious; 2) prevention of operational errors resulting from misconfiguration within the trusted zone.

- o MPLS-TP MUST support the physical and logical separation of data plane from the control plane and management plane. That is, if the control plane or/and management plane are attacked and cannot function normally, the data plane should continue to forward packets without being impacted.
- o MPLS-TP MUST support static provisioning of MPLS-TP LSP and PW with or without NMS/OSS, without using control protocols. This is particularly important in the case of security model 2(a)

(Figure 3) and security model 2(b) (Figure 4) where some or all T-PEs are not in the trusted zone, and in the inter-provider cases in security model 2(c) (Figure 5) when the connecting S-PE is in the untrusted zone.

- o MPLS-TP MUST support non-IP path options in addition to IP loopback option. Non-IP path options when used in security model 2 (Section 2.2) may help to lower the potential risk of attack on the S-PE/T-PE in the trusted zone.
- o MPLS-TP MUST support authentication of any control protocol used for an MPLS-TP network, as well as for MPLS-TP network to dynamic MPLS network inter-connection.
- o MPLS-TP MUST support mechanisms to prevent Denial of Service (DOS) attacks via any in-band OAM or G-ACh/GAL.
- o MPLS-TP MUST support hiding of the Service Provider infrastructure for all reference models regardless of whether the network(s) are using static configuration or a dynamic control plane.
- o Security management requirements from [RFC5951]:
 - * MPLS-TP MUST support management communication channel (MCC) security.
 - * Secure communication channels MUST be supported for all network traffic and protocols used to support management functions. This MUST include protocols used for configuration, monitoring, configuration backup, logging, time synchronization, authentication, and routing.
 - * The MCC MUST support application protocols that provide confidentiality and data integrity protection.
 - * The MCC MUST support the use of open cryptographic algorithms [RFC3871].
 - * The MCC MUST support authentication to ensure that management connectivity and activity is only from authenticated entities.
 - * The MCC MUST support port access control.
 - * Distributed Denial of Service: It is possible to lessen the impact and potential for DoS and DDoS by using secure protocols, turning off unnecessary processes, logging and monitoring, and ingress filtering. [RFC4732] provides background on DOS in the context of the Internet.

- o MPLS-TP MUST provide protection from operational error. Due to the extensive use of static provisioning with or without NMS and OSS, the prevention of configuration errors should be addressed as major security requirements.

4. Security Threats

This section discusses the various network security threats that may endanger MPLS-TP networks. The discussion is limited to those threats that are unique to MPLS-TP networks or that affect MPLS-TP networks in unique ways.

A successful attack on a particular MPLS-TP network or on a SP's MPLS-TP infrastructure may cause one or more of the following ill effects:

1. Observation (including traffic pattern analysis), modification, or deletion of a provider's or user's data, as well as replay or insertion of non-authentic data into a provider's or user's data stream. These types of attacks apply to MPLS-TP traffic regardless of how the LSP or PW is set up in a similar way to how they apply to MPLS traffic regardless how the LSP is set up.
2. Attacks on GAL label, BFD messages:
 1. GAL label or BFD label manipulation: including insertion of false label or messages, or modification, or removal the GAL labels or messages by attackers.
 2. DOS attack through in-band OAM G-ACH/GAL, and BFD messages.
3. Disruption of a provider's and/or user's connectivity, or degradation of a provider's service quality.
 1. Provider connectivity attacks:
 - + In the case of NMS is used for LSP set-up, the attacks would be through the attack of NMS.
 - + In the case of dynamic is used for dynamic provisioning, the attack would be on dynamic control plane. Most aspects are addressed in [RFC5920].
 2. User connectivity attack. This would be similar as PE/CE access attack in typical MPLS networks, addressed in [RFC5920].

4. Probing a provider's network to determine its configuration, capacity, or usage. These types of attack can happen through NMS attacks in the case of static provisioning, or through control plane attacks as in dynamic MPLS networks. It can also be combined attacks.

It is useful to consider that threats, whether malicious or accidental, may come from different categories of sources. For example they may come from:

- o Other users whose services are provided by the same MPLS-TP core.
- o The MPLS-TP SP or persons working for it.
- o Other persons who obtain physical access to a MPLS-TP SP's site.
- o Other persons who use social engineering methods to influence the behavior of a SP's personnel.
- o Users of the MPLS-TP network itself.
- o Others, e.g., attackers from the other sources, Internet if connected.
- o Other SPs in the case of MPLS-TP Inter-provider connection. The provider may or may not be using MPLS-TP.
- o Those who create, deliver, install, and maintain software for network equipment.

Given that security is generally a tradeoff between expense and risk, it is also useful to consider the likelihood of different attacks occurring. There is at least a perceived difference in the likelihood of most types of attacks being successfully mounted in different environments, such as:

- o A MPLS-TP network inter-connecting with another provider's core
- o A MPLS-TP configuration transiting the public Internet

Most types of attacks become easier to mount and hence more likely as the shared infrastructure via which service is provided expands from a single SP to multiple cooperating SPs to the global Internet. Attacks that may not be of sufficient likeliness to warrant concern in a closely controlled environment often merit defensive measures in broader, more open environments. In closed communities, it is often practical to deal with misbehavior after the fact: an employee can be disciplined, for example.

The following sections discuss specific types of exploits that threaten MPLS-TP networks.

4.1. Attacks on the Control Plane

- o MPLS-TP LSP creation by an unauthorized element
- o LSP message interception
- o Attacks on G-Ach
- o Attacks against LDP
- o Attacks against RSVP-TE
- o Attacks against GMPLS
- o Denial of Service Attacks on the Network Infrastructure
- o Attacks on the SP's MPLS/GMPLS Equipment via Management Interfaces
- o Social Engineering Attacks on the SP's Infrastructure
- o Cross-Connection of Traffic between Users
- o Attacks against Routing Protocols
- o Other Attacks on Control Traffic

4.2. Attacks on the Data Plane

This category encompasses attacks on the provider's or end user's data. Note that from the MPLS-TP network end user's point of view, some of this might be control plane traffic, e.g. routing protocols running from user site A to user site B via IP or non-IP connections, which may be some type of VPN.

- o Unauthorized Observation of Data Traffic
- o Modification of Data Traffic
- o Insertion of Inauthentic Data Traffic: Spoofing and Replay
- o Unauthorized Deletion of Data Traffic
- o Unauthorized Traffic Pattern Analysis

- o Denial of Service Attacks
- o Misconnection

5. Defensive Techniques for MPLS-TP Networks

The defensive techniques discussed in this document are intended to describe methods by which some security threats can be addressed. They are not intended as requirements for all MPLS-TP implementations. The MPLS-TP provider should determine the applicability of these techniques to the provider's specific service offerings, and the end user may wish to assess the value of these techniques to the user's service requirements. The operational environment determines the security requirements. Therefore, protocol designers need to provide a full set of security services, which can be used where appropriate.

The techniques discussed here include encryption, authentication, filtering, firewalls, access control, isolation, aggregation, and others.

5.1. Authentication

To prevent security issues arising from some DoS attacks or from malicious or accidental misconfiguration, it is critical that devices in the MPLS-TP should only accept connections or control messages from valid sources. Authentication refers to methods to ensure that message sources are properly identified by the MPLS-TP devices with which they communicate. This section focuses on identifying the scenarios in which sender authentication is required and recommends authentication mechanisms for these scenarios.

5.1.1. Management System Authentication

Management system authentication includes the authentication of a PE to a centrally-managed network management or directory server when directory-based "auto-discovery" is used. It also includes authentication of a CE to the configuration server, when a configuration server system is used.

Authentication should be bi-directional, including PE or CE to configuration server authentication for PE or CE to be certain it is communicating with the right server.

5.1.2. Peer-to-Peer Authentication

Peer-to-peer authentication includes peer authentication for network control protocols and other peer authentication (i.e., authentication of one IPsec security gateway by another).

Authentication should be bi-directional, including S-PE, T-PE, PE or CE to configuration server authentication for PE or CE to be certain it is communicating with the right server.

5.1.3. Cryptographic Techniques for Authenticating Identity

Cryptographic techniques offer several mechanisms for authenticating the identity of devices or individuals. These include the use of shared secret keys, one-time keys generated by accessory devices or software, user-ID and password pairs, and a range of public-private key systems. Another approach is to use a hierarchical Certification Authority system to provide digital certificates.

5.2. Access Control Techniques

Most of the security issues related to management interfaces can be addressed through the use of authentication techniques as described in the section on authentication. However, additional security may be provided by controlling access to management interfaces in other ways.

The Optical Internetworking Forum has done relevant work on protecting such interfaces with TLS, SSH, Kerberos, IPsec, WSS, etc. See Security for Management Interfaces to Network Elements [OIF-SMI-01.0], and Addendum to the Security for Management Interfaces to Network Elements [OIF-SMI-02.1]. See also the work in the ISMS WG.

Management interfaces, especially console ports on MPLS-TP devices, may be configured so they are only accessible out-of-band, through a system which is physically or logically separated from the rest of the MPLS-TP infrastructure.

Where management interfaces are accessible in-band within the MPLS-TP domain, filtering or firewalling techniques can be used to restrict unauthorized in-band traffic from having access to management interfaces. Depending on device capabilities, these filtering or firewalling techniques can be configured either on other devices through which the traffic might pass, or on the individual MPLS-TP devices themselves.

5.3. Use of Isolated Infrastructure

One way to protect the infrastructure used for support of MPLS-TP is to separate the resources for support of MPLS-TP services from the resources used for other purposes.

5.4. Use of Aggregated Infrastructure

In general, it is not feasible to use a completely separate set of resources for support of each service. In fact, one of the main reasons for MPLS-TP enabled services is to allow sharing of resources between multiple services and multiple users. Thus, even if certain services use a separate network from Internet services, nonetheless there will still be multiple MPLS-TP users sharing the same network resources.

In general, the use of aggregated infrastructure allows the service provider to benefit from stochastic multiplexing of multiple bursty flows, and also may in some cases thwart traffic pattern analysis by combining the data from multiple users. However, service providers must minimize security risks introduced from any individual service or individual users.

5.5. Service Provider Quality Control Processes

5.6. Verification of Connectivity

In order to protect against deliberate or accidental misconnection, mechanisms can be put in place to verify both end-to-end connectivity and hop-by-hop resources. These mechanisms can trace the routes of LSPs in both the control plane and the data plane.

6. Monitoring, Detection, and Reporting of Security Attacks

MPLS-TP network and service may be subject to attacks from a variety of security threats. Many threats are described in the Security Requirements (Section 3) Section of this document. Many of the defensive techniques described in this document and elsewhere provide significant levels of protection from a variety of threats. However, in addition to employing defensive techniques silently to protect against attacks, MPLS-TP services can also add value for both providers and customers by implementing security monitoring systems to detect and report on any security attacks, regardless of whether the attacks are effective.

Attackers often begin by probing and analyzing defenses, so systems that can detect and properly report these early stages of attacks can

provide significant benefits.

Information concerning attack incidents, especially if available quickly, can be useful in defending against further attacks. It can be used to help identify attackers or their specific targets at an early stage. This knowledge about attackers and targets can be used to strengthen defenses against specific attacks or attackers, or to improve the defenses for specific targets on an as-needed basis. Information collected on attacks may also be useful in identifying and developing defenses against novel attack types.

7. Security Considerations

Security considerations constitute the sole subject of this memo and hence are discussed throughout.

The document describes a variety of defensive techniques that may be used to counter the suspected threats. All of the techniques presented involve mature and widely implemented technologies that are practical to implement.

The document evaluates MPLS-TP security requirements from a customer's perspective as well as from a service provider's perspective. These sections re-evaluate the identified threats from the perspectives of the various stakeholders and are meant to assist equipment vendors and service providers, who must ultimately decide what threats to protect against in any given configuration or service offering.

8. IANA Considerations

This document contains no new IANA considerations.

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.
- [RFC3871] Jones, G., "Operational Security Requirements for Large Internet Service Provider (ISP) IP Network Infrastructure", RFC 3871, September 2004.
- [RFC4732] Handley, M., Rescorla, E., and IAB, "Internet Denial-of-

Service Considerations", RFC 4732, December 2006.

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

[RFC5951] Lam, K., Mansfield, S., and E. Gray, "Network Management Requirements for MPLS-based Transport Networks", RFC 5951, September 2010.

9.2. Informative References

[OIF-SMI-01.0]
Optical Internetworking Forum, "Security for Management Interfaces to Network Elements", OIF OIF-SMI-01.0, Sept 2003.

[OIF-SMI-02.1]
Optical Internetworking Forum, "Addendum to the Security for Management Interfaces to Network Elements", OIF OIF-SMI-02.1, March 2006.

[RFC3631] Bellovin, S., Schiller, J., and C. Kaufman, "Security Mechanisms for the Internet", RFC 3631, December 2003.

[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.

[opsec-efforts]
"Security Best Practices Efforts and Documents", IETF draft-ietf-opsec-efforts-08.txt, June 2008.

Authors' Addresses

Luyuan Fang (editor)
Cisco Systems Inc.
111 Wood Ave. South
Iselin, NJ 08830
US

Email: lufang@cisco.com

Ben Niven-Jenkins (editor)
Velocix
326 Cambridge Science Park
Milton Road
Cambridge CB4 0WG
UK

Email: ben@niven-jenkins.co.uk

Scott Mansfield (editor)
Ericsson
300 Holger Way
San Jose, CA 95134
US

Email: scott.mansfield@ericsson.com

Raymond Zhang
British Telecom
BT Center
81 Newgate Street
London EC1A 7AJ
Uk

Email: raymond.zhang@bt.com

Nabil Bitar
Verizon
40 Sylvan Road
Waltham, MA 02145
US

Email: nabil.bitar@verizon.com

Masahiro Daikoku
KDDI Corporation
3-11-11 Iidabashi, Chiyodaku
Tokyo
Japan

Email: ms-daikoku@kddi.com

Lei Wang
Telenor
Telenor Norway
Office Snaroyveien
1331 Fornedbu
Norway

Email: lei.wang@telenor.com

Henry Yu
TW Telecom
10475 Park Meadow Drive
Littleton, CO 80124
US

Email: henry.yu@twtelecom.com

Network Working Group
INTERNET-DRAFT
Intended Status: Standards Track
Expires: December 17, 2011

M.Venkatesan
Kannan KV Sampath
Aricent
Sam K. Aldrin
Huawei Technologies
Thomas D. Nadeau
CA Technologies

June 17, 2011

MPLS-TP Traffic Engineering (TE) Management Information Base (MIB)
draft-ietf-mpls-tp-te-mib-00.txt

Abstract

This memo defines a portion of the Management Information Base (MIB) for use with network management protocols in the Internet community. In particular, it describes managed objects of Tunnels, Identifiers, Label Switch Router and Textual conventions for Multiprotocol Label Switching (MPLS) based Transport Profile (TP).

Status of this Memo

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

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

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

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

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

This Internet-Draft will expire on December 17, 2011.

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.

Table of Contents

1	Introduction	3
2.	The Internet-Standard Management Framework	3
3.	Overview	3
3.1	Conventions used in this document	3
3.2	Terminology	3
3.3	Acronyms	3
4.	Motivations	4
5.	Feature List	4
6.	Brief description of MIB Objects	4
6.1.	mplsNodeConfigTable	5
6.2.	mplsNodeIpMapTable	5
6.3.	mplsNodeIccMapTable	6
6.4.	mplsTunnelExtTable	6
7.	MIB Module Interdependencies	6
8.	Dependencies between MIB Module Tables	8
9.	Example of MPLS-TP tunnel setup	8
10.	MPLS Textual Convention Extension MIB definitions	13
11.	MPLS Identifier MIB definitions	16
12.	MPLS LSR Extension MIB definitions	20
13.	MPLS Tunnel Extension MIB definitions	24
14.	Security Consideration	36
15.	IANA Considerations	37
16.	References	37
16.1	Normative References	37
16.2	Informative References	38
17.	Acknowledgments	38
18.	Authors' Addresses	38

1 Introduction

This memo defines a portion of the Management Information Base (MIB) for use with network management protocols in the Internet community. In particular, it describes managed objects of Tunnels, Identifiers, Label Switch Router and Textual conventions for Multiprotocol Label Switching (MPLS) based Transport Profile (TP).

This MIB module should be used in conjunction with the MPLS traffic Engineering MIB [RFC3812] and companion document MPLS Label Switch Router MIB [RFC3813] for MPLS based traffic engineering configuration and management.

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 BCP 14, RFC2119.

2. The Internet-Standard Management Framework

For a detailed overview of the documents that describe the current Internet-Standard Management Framework, please refer to section 7 of RFC3410 [RFC3410].

Managed objects are accessed via a virtual information store, termed the Management Information Base or MIB. MIB objects are generally accessed through the Simple Network Management Protocol (SNMP). Objects in the MIB are defined using the mechanisms defined in the Structure of Management Information (SMI). This memo specifies a MIB module that is compliant to the SMIv2, which is described in STD 58, RFC2578, STD 58, RFC2579 and STD58, RFC2580.

3. Overview

3.1 Conventions used in this document

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

3.2 Terminology

This document uses terminology from the MPLS architecture document [RFC3031], MPLS Traffic Engineering Management information [RFC3812], MPLS Label Switch Router MIB [RFC3813] and MPLS-TP Identifiers document [TPIDS].

3.3 Acronyms

GMPLS: Generalized Multi-Protocol Label Switching
ICC: ITU Carrier Code
IP: Internet Protocol
LSP: Label Switching Path
LSR: Label Switching Router
MIB: Management Information Base
MPLS: Multi-Protocol Label Switching
MPLS-TP: MPLS Transport Profile
OSPF: Open Shortest Path First
PW: Pseudowire
TE: Traffic Engineering
TP: Transport Profile

4. Motivations

The existing MPLS TE [RFC3812] and GMPLS MIBs [RFC4802] do not support the transport network requirements of NON-IP based management and static bidirectional tunnels.

5. Feature List

The MPLS transport profile MIB module is designed to satisfy the following requirements and constraints:

The MIB module supports point-to-point, co-routed bi-directional associated bi-directional tunnels.

- The MPLS tunnels need not be interfaces, but it is possible to configure a TP tunnel as an interface.
- The `mplsTunnelTable` [RFC3812] to be also used for MPLS-TP tunnels
- The `mplsTunnelTable` is extended to support MPLS-TP specific objects.
- A node configuration table (`mplsNodeConfigTable`) is used to translate the `Global_Node_ID` or `ICC` to the local identifier in order to index `mplsTunnelTable`.
- The MIB module supports persistent, as well as non-persistent tunnels.

6. Brief description of MIB Objects

The objects described in this section support the functionality described in documents [RFC5654] and [TPIDS]. The tables support

both IP compatible and ICC based tunnel configurations.

6.1. mplsNodeConfigTable

The mplsNodeConfigTable is used to assign a local identifier for a given ICC or Global_Node_ID combination as defined in [TPIDS]. An ICC is a string of one to six characters, each character being either alphabetic (i.e. A-Z) or numeric (i.e. 0-9) characters. Alphabetic characters in the ICC should be represented with upper case letters. In the IP compatible mode, Global_Node_ID, is used to uniquely identify a node.

Each ICC or Global_Node_ID contains one unique entry in the table representing a node. Every node is assigned a local identifier within a range of 0 to 16777215. This local identifier is used for indexing into mplsTunnelTable as mplsTunnelIngressLSRId and mplsTunnelEgressLSRId.

For IP compatible environment, MPLS-TP tunnel is indexed by Tunnel Index, Tunnel Instance, Source Global_ID, Source Node_ID, Destination Global_ID and Destination Node_ID.

For ICC based environment, MPLS-TP tunnel is indexed by Tunnel Index, Tunnel Instance, Source ICC and Destination ICC.

As mplsTunnelTable is indexed by mplsTunnelIndex, mplsTunnelInstance, mplsTunnelIngressLSRId, and mplsTunnelEgressLSRId, the MPLS-TP tunnel identifiers cannot be used directly.

The mplsNodeConfigTable will be used to store an entry for ICC or Global_Node_ID with a local identifier to be used as LSR ID in mplsTunnelTable. As the regular TE tunnels use IP address as LSR ID, the local identifier should be below the first valid IP address, which is 16777216[1.0.0.0].

6.2. mplsNodeIpMapTable

The read-only mplsNodeIpMapTable is used to query the local identifier assigned and stored in mplsNodeConfigTable for a given Global_Node_ID. In order to query the local identifier, in the IP compatible mode, this table is indexed with Global_Node_ID. In the IP compatible mode for a TP tunnel, Global_Node_ID is used.

A separate query is made to get the local identifier of both Ingress and Egress Global_Node_ID identifiers. These local

identifiers are used as `mplsTunnelIngressLSRId` and `mplsTunnelEgressLSRId`, while indexing `mplsTunnelTable`.

6.3. `mplsNodeIccMapTable`

The read-only `mplsNodeIccMapTable` is used to query the local identifier assigned and stored in the `mplsNodeConfigTable` for a given ICC.

A separate query is made to get the local identifier of both Ingress and Egress ICC. These local identifiers are used as `mplsTunnelIngressLSRId` and `mplsTunnelEgressLSRId`, while indexing `mplsTunnelTable`.

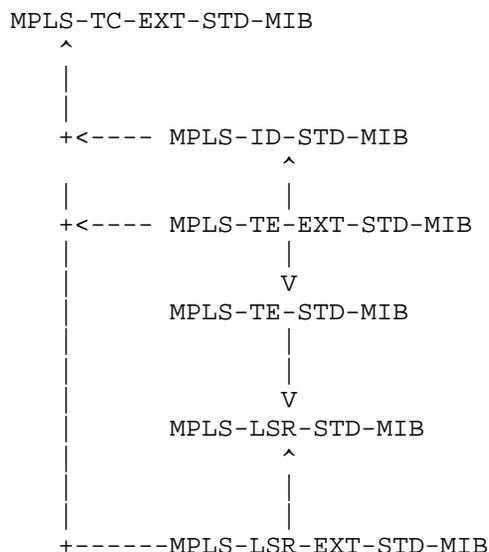
6.4. `mplsTunnelExtTable`

`mplsTunnelExtTable` extends the `mplsTunnelTable` to add MPLS-TP tunnel specific additional objects. All the additional attributes specific to TP tunnel are contained in this extended table and could be accessed with the `mplsTunnelTable` indices.

7. MIB Module Interdependencies

This section provides an overview of the relationship between the MPLS-TP TE MIB module and other MPLS MIB modules.

The arrows in the following diagram show a 'depends on' relationship. A relationship "MIB module A depends on MIB module B" means 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 new MPLS extension MIB modules depend on MPLS-TC-EXT-STD-MIB.
- MPLS-TE-STD-MIB [RFC3812] contains references to objects in MPLS-ID-STD-MIB.
- MPLS-TE-EXT-STD-MIB contains references to objects in MPLS-TE-STD-MIB [RFC3812].
- MPLS-LSR-EXT-STD-MIB contains references to objects in MPLS-LSR-STD-MIB [RFC3813].

MPLS-TE-STD-MIB [RFC 3812] is extended by MPLS-TE-EXT-STD-MIB mib module for associating the reverse direction tunnel information.

Note that the nature of the 'extends' relationship is a sparse augmentation so that the entry in the mplsTunnelExtTable has the same index values as the in the mplsTunnelTable.

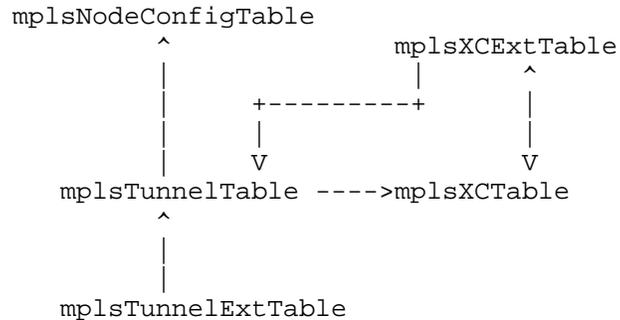
MPLS-LSR-STD-MIB [RFC 3813] is extended by MPLS-LSR-EXT-STD-MIB mib module for pointing back to the tunnel entry for easy tunnel access from XC entry.

Note that the nature of the 'extends' relationship

is a sparse augmentation so that the entry in the `mplsXCExtTable` has the same index values as the in the `mplsXCTable`.

8. Dependencies between MIB Module Tables

The tables in MPLS-TE-EXT-STD-MIB are related as shown on the diagram below. The arrows indicate a reference from one table to another.



An existing `mplsTunnelTable` uses the new `mplsNodeConfigTable` table to map the `Global_Node_ID` and/or `ICC` with the local number in order to accommodate in the existing tunnel table's ingress/egress LSR-id.

New `mplsTunnelExtTable` table provides the reverse direction LSP information for the existing tunnel table in order to achieve bidirectional LSPs.

`mplsXCExtTable` is extended from `mplsLsrXCTable` to provide backward reference to tunnel entry.

9. Example of MPLS-TP tunnel setup

In this section, we provide an example of the IP based MPLS-TP co-routed bidirectional tunnel setup. This example provides the usage of MPLS-TP Tunnel MIB along with the extended new MIB modules introduced in this document.

Do note that a MPLS-TP tunnel could be setup statically as well as signaled via control plane. This example considers configuration on a head-end LSR to setup a static MPLS-TP tunnel. Only relevant objects which are applicable for MPLS-TP tunnel are illustrated here.

In `mplsNodeConfigTable`:

```

{
  -- Non-IP Ingress LSR-Id (Index to the table)
  mplsNodeConfigLocalId = 1,

```

```

mplsNodeConfigGlobalId      = 1234,
mplsNodeConfigNodeId        = 10,
-- Mandatory parameters needed to activate the row go here
mplsNodeConfigRowStatus     = createAndGo (4)

-- Non-IP Egress LSR-Id (Index to the table)
mplsNodeConfigLocalId       = 2,
mplsNodeConfigGlobalId      = 1234,
mplsNodeConfigNodeId        = 20,
-- Mandatory parameters needed to activate the row go here
mplsNodeConfigRowStatus     = createAndGo (4)
}

```

This will create an entry in the mplsNodeConfigTable for a Global_Node_ID. A separate entry is made for both Ingress LSR and Egress LSR.

The following read-only mplsNodeIpMapTable table is populated automatically upon creating an entry in mplsNodeConfigTable and this table is used to retrieve the local identifier for the given Global_Node_ID.

In mplsNodeIpMapTable:

```

{
-- Global_ID (Index to the table)
mplsNodeIpMapGlobalId      = 1234,
-- Node Identifier (Index to the table)
mplsNodeIpMapNodeId        = 10,
mplsNodeIpMapLocalId       = 1

-- Global_ID (Index to the table)
mplsNodeIpMapGlobalId      = 1234,
-- Node Identifier (Index to the table)
mplsNodeIpMapNodeId        = 20,
mplsNodeIpMapLocalId       = 2
}

```

The following denotes the configured tunnel "head" entry:

In mplsTunnelTable:

```

{
mplsTunnelIndex            = 1,
mplsTunnelInstance         = 1,
-- Local map number created in mplsNodeConfigTable for Ingress
  LSR-Id
mplsTunnelIngressLSRId     = 1,

```

```

-- Local map number created in mplsNodeConfigTable for Egress
  LSR-Id
  mplsTunnelEgressLSRId      = 2,
  mplsTunnelName             = "TP forward LSP",
  mplsTunnelDescr           = "East to West",
  mplsTunnelIsIf            = true (1),
-- RowPointer MUST point to the first accessible column
  mplsTunnelXCPointer       =
                                mplsXCLspId.4.0.0.0.1.1.0.4.0.0.0.12,
  mplsTunnelSignallingProto = none (1),
  mplsTunnelSetupPrio       = 0,
  mplsTunnelHoldingPrio    = 0,
  mplsTunnelSessionAttributes = 0,
  mplsTunnelLocalProtectInUse = false (0),
-- RowPointer MUST point to the first accessible column
  mplsTunnelResourcePointer = mplsTunnelResourceMaxRate.5,
  mplsTunnelInstancePriority = 1,
  mplsTunnelHopTableIndex   = 1,
  mplsTunnelIncludeAnyAffinity = 0,
  mplsTunnelIncludeAllAffinity = 0,
  mplsTunnelExcludeAnyAffinity = 0,
  mplsTunnelRole            = head (1),
-- Mandatory parameters needed to activate the row go here
  mplsTunnelRowStatus       = createAndGo (4)
}

```

In mplsTunnelTable:

```

{
  mplsTunnelIndex           = 1,
  mplsTunnelInstance       = 2,
-- Local map number created in mplsNodeConfigTable for Ingress
  LSR-Id
  mplsTunnelIngressLSRId   = 1,
-- Local map number created in mplsNodeConfigTable for Egress
  LSR-Id
  mplsTunnelEgressLSRId    = 2,
  mplsTunnelName           = "TP reverse LSP",
  mplsTunnelDescr         = "West to East",
  mplsTunnelIsIf          = true (1),
-- RowPointer MUST point to the first accessible column
  mplsTunnelXCPointer      =
                                mplsXCLspId.4.0.0.0.1.4.0.0.0.16.1.0,
  mplsTunnelSignallingProto = none (1),
  mplsTunnelSetupPrio      = 0,
  mplsTunnelHoldingPrio    = 0,
  mplsTunnelSessionAttributes = 0,
  mplsTunnelLocalProtectInUse = false (0),
}

```

```

-- RowPointer MUST point to the first accessible column
mplsTunnelResourcePointer    = mplsTunnelResourceMaxRate.5,
mplsTunnelInstancePriority    = 1,
mplsTunnelHopTableIndex      = 1,
mplsTunnelIncludeAnyAffinity = 0,
mplsTunnelIncludeAllAffinity = 0,
mplsTunnelExcludeAnyAffinity = 0,
mplsTunnelRole                = head (1),
-- Mandatory parameters needed to activate the row go here
mplsTunnelRowStatus          = createAndGo (4)
}

```

Now the TP specific Tunnel parameters are configured in the extended Tunnel table

In mplsTunnelExtTable:

```

{
  Index = same as one used for mplsTunnelTable,
  -- As per [TPIDS] LSP_ID is defined as follows,
  -- For corouted bidirectional tunnel
  -- LSP_ID => East-Global_Node_ID::East-Tunnel_Num::
  --           West-Global_Node_ID::West-Tunnel_Num::LSP_Num
  -- LSP_ID of this tunnel: 1234_10::1::1234_20::1::0
  -- Where,
  -- LSP_Num - 0 indicates the configured head end tunnel.

  -- West tunnel number is assigned in the destination
  -- tunnel index,
  -- single LSP number is common for both forward and reverse
  -- directions, as the single tunnel head entry originates
  -- both the forward and reverse LSPs.
  -- mplsTunnelExtDestTnlIndex = West-Tunnel_Num
  -- mplsTunnelExtDestTnlLspIndex = LSP_Num

  mplsTunnelExtDestTnlIndex    = 1,
  mplsTunnelExtDestTnlLspIndex = 0

  -- For associated bidirectional tunnel
  -- LSP_ID => East-Global_Node_ID::East-Tunnel_Num::
  --           East-LSP_Num::West-Global_Node_ID::
  --           West-Tunnel_Num::West-LSP_Num
  -- West tunnel number is assigned in the destination
  -- tunnel index, since the head end tunnel is different for
  -- both the forward and reverse direction LSPs,
  -- Destination LSP index points the reverse direction LSP
  -- in a different tunnel.
  -- mplsTunnelExtDestTnlIndex = West-Tunnel_Num
  -- mplsTunnelExtDestTnlLspIndex = West-LSP_Num
}

```

```
}

```

We must next create the appropriate in-segment and out-segment entries. These are done in [RFC3813] using the `mplsInSegmentTable` and `mplsOutSegmentTable`.

For the forward direction.

```
In mplsOutSegmentTable:
{
  mplsOutSegmentIndex      = 0x00000012,
  mplsOutSegmentInterface  = 13, -- outgoing interface
  mplsOutSegmentPushTopLabel = true(1),
  mplsOutSegmentTopLabel   = 22, -- outgoing label

  -- RowPointer MUST point to the first accessible column.
  mplsOutSegmentTrafficParamPtr = 0.0,
  mplsOutSegmentRowStatus      = createAndGo (4)
}
```

For the reverse direction.

```
In mplsInSegmentTable:
{
  mplsInSegmentIndex      = 0x00000016
  mplsInSegmentLabel      = 21, -- incoming label
  mplsInSegmentNPop       = 1,
  mplsInSegmentInterface  = 13, -- incoming interface

  -- RowPointer MUST point to the first accessible column.
  mplsInSegmentTrafficParamPtr = 0.0,
  mplsInSegmentRowStatus      = createAndGo (4)
}
```

Next, two cross-connect entries are created in the `mplsXCTable` of the MPLS-LSR-STD-MIB [RFC3813], thereby associating the newly created segments together.

```
In mplsXCTable:
{
  mplsXCIndex              = 0x01,
  mplsXCInSegmentIndex    = 0x00000000,
  mplsXCOutSegmentIndex   = 0x00000012,
  mplsXCLspId             = 0x0102 -- unique ID
  -- only a single outgoing label
  mplsXCLabelStackIndex   = 0x00,
  mplsXCRowStatus         = createAndGo(4)
}
```

```

}

In mplsXCTable:
{
  mplsXCIndex                = 0x01,
  mplsXCInSegmentIndex      = 0x00000016,
  mplsXCOutSegmentIndex     = 0x00000000,
  mplsXCCLspID              = 0x0102 -- unique ID
  -- only a single outgoing label
  mplsXCLabelStackIndex     = 0x00,
  mplsXCRowStatus           = createAndGo(4)
}

```

This table entry is extended by entry in the mplsXCExtTable. Note that the nature of the 'extends' relationship is a sparse augmentation so that the entry in the mplsXCExtTable has the same index values as the entry in the mplsXCTable.

First for the forward direction:

```

In mplsXCExtTable
{
  -- Back pointer from XC table to Tunnel table
  mplsXCExtTunnelPointer = mplsTunnelName.1.1.1.2
}

```

Next for the reverse direction:

```

In mplsXCExtTable
{
  -- Back pointer from XC table to Tunnel table
  mplsXCExtTunnelPointer = mplsTunnelName.1.2.1.2
}

```

10. MPLS Textual Convention Extension MIB definitions

```
MPLS-TC-EXT-STD-MIB DEFINITIONS ::= BEGIN
```

```
IMPORTS
```

```
  MODULE-IDENTITY, Unsigned32
    FROM SNMPv2-SMI -- [RFC2578]
```

```
  TEXTUAL-CONVENTION
    FROM SNMPv2-TC -- [RFC2579]
```

```
  mplsStdMIB
    FROM MPLS-TC-STD-MIB -- [RFC3811]
```

;

mplsTcExtStdMIB MODULE-IDENTITY

LAST-UPDATED

"201106160000Z" -- June 16, 2011

ORGANIZATION

"Multiprotocol Label Switching (MPLS) Working Group"

CONTACT-INFO

"

Venkatesan Mahalingam
Aricent,
India

Email: venkatesan.mahalingam@aricent.com

Kannan KV Sampath
Aricent,
India

Email: Kannan.Sampath@aricent.com

Sam Aldrin
Huawei Technologies
2330 Central Express Way,
Santa Clara, CA 95051, USA

Email: aldrin.ietf@gmail.com

Thomas D. Nadeau
CA Technologies
273 Corporate Drive, Portsmouth, NH, USA

Email: thomas.nadeau@ca.com

"

DESCRIPTION

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

This MIB module contains Textual Conventions for
MPLS based transport networks."

-- Revision history.

REVISION

"201106160000Z" -- June 16, 2011

DESCRIPTION

"MPLS Textual Convention Extensions"

::= { mplsStdMIB xxx } -- xxx to be replaced with correct value

MplsGlobalId ::= TEXTUAL-CONVENTION

STATUS current
DESCRIPTION
"This object contains the Textual Convention of IP based operator unique identifier (Global_ID), the Global_ID can contain the 2-octet or 4-octet value of the operator's Autonomous System Number (ASN).

It is expected that the Global_ID will be derived from the globally unique ASN of the autonomous system hosting the PEs containing the actual AIIIs.
The presence of a Global_ID based on the operator's ASN ensures that the AII will be globally unique.

When the Global_ID is derived from a 2-octet AS number, the two high-order octets of this 4-octet identifier MUST be set to zero.
Further ASN 0 is reserved. A Global_ID of zero means that no Global_ID is present. Note that a Global_ID of zero is limited to entities contained within a single operator and MUST NOT be used across an NNI.
A non-zero Global_ID MUST be derived from an ASN owned by the operator."
SYNTAX OCTET STRING (SIZE (4))

MplsNodeId ::= TEXTUAL-CONVENTION
DISPLAY-HINT "d"
STATUS current
DESCRIPTION
"The Node_ID is assigned within the scope of the Global_ID. The value 0(or 0.0.0.0 in dotted decimal notation) is reserved and MUST NOT be used.

When IPv4 addresses are in use, the value of this object can be derived from the LSR's /32 IPv4 loop back address.

Note that, when IP reach ability is not needed, the 32-bit Node_ID is not required to have any association with the IPv4 address space."
SYNTAX Unsigned32

MplsIccId ::= TEXTUAL-CONVENTION
STATUS current
DESCRIPTION
"The ICC is a string of one to six characters, each character being either alphabetic (i.e. A-Z) or numeric (i.e. 0-9) characters.
Alphabetic characters in the ICC SHOULD be represented

with upper case letters."
 SYNTAX OCTET STRING (SIZE (1..6))

MplsLocalId ::= TEXTUAL-CONVENTION
 DISPLAY-HINT "d"
 STATUS current
 DESCRIPTION
 "This textual convention is used in accommodating the bigger
 size Global_Node_ID and/or ICC with lower size LSR identifier
 in order to index the mplsTunnelTable.

 The Local Identifier is configured between 1 and 16777215,
 as valid IP address range starts from 16777216 (01.00.00.00).
 This range is chosen to identify the mplsTunnelTable's
 Ingress/Egress LSR-id is IP address or Local identifier,
 if the configured range is not IP address, administrator is
 expected to retrieve the complete information (Global_Node_ID
 or ICC) from mplsNodeConfigTable. This way, existing
 mplsTunnelTable is reused for bidirectional tunnel extensions
 for MPLS based transport networks.

This Local Identifier allows the administrator to assign
 a unique identifier to map Global_Node_ID and/or ICC."
 SYNTAX Unsigned32(1..16777215)

-- MPLS-TC-EXT-STD-MIB module ends
 END

11. MPLS Identifier MIB definitions

```
MPLS-ID-STD-MIB DEFINITIONS ::= BEGIN

IMPORTS
  MODULE-IDENTITY, OBJECT-TYPE, NOTIFICATION-TYPE
    FROM SNMPv2-SMI -- [RFC2578]
  MODULE-COMPLIANCE, OBJECT-GROUP, NOTIFICATION-GROUP
    FROM SNMPv2-CONF -- [RFC2580]
  mplsStdMIB
    FROM MPLS-TC-STD-MIB -- [RFC3811]
  MplsGlobalId, MplsIccId, MplsNodeId
    FROM MPLS-TC-EXT-STD-MIB
;

mplsIdStdMIB MODULE-IDENTITY
  LAST-UPDATED
    "201106160000Z" -- June 16, 2011
  ORGANIZATION
    "Multiprotocol Label Switching (MPLS) Working Group"
```

CONTACT-INFO

"

Venkatesan Mahalingam
Aricent,
India

Email: venkatesan.mahalingam@aricent.com

Kannan KV Sampath
Aricent,
India

Email: Kannan.Sampath@aricent.com

Sam Aldrin
Huawei Technologies
2330 Central Express Way,
Santa Clara, CA 95051, USA

Email: aldrin.ietf@gmail.com

Thomas D. Nadeau
CA Technologies
273 Corporate Drive, Portsmouth, NH, USA

Email: thomas.nadeau@ca.com

"

DESCRIPTION

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

This MIB module contains generic object definitions for
MPLS Traffic Engineering in transport networks."

-- Revision history.

REVISION

"201106160000Z" -- June 16, 2011

DESCRIPTION

"MPLS identifiers mib object extension"

::= { mplsStdMIB xxx } -- xxx to be replaced with correct value

-- traps

mplsIdNotifications OBJECT IDENTIFIER ::= { mplsIdStdMIB 0 }

-- tables, scalars

mplsIdObjects OBJECT IDENTIFIER ::= { mplsIdStdMIB 1 }

-- conformance

mplsIdConformance OBJECT IDENTIFIER ::= { mplsIdStdMIB 2 }

```
-- MPLS common objects

mplsGlobalId OBJECT-TYPE
    SYNTAX      MplsGlobalId
    MAX-ACCESS  read-write
    STATUS      current
    DESCRIPTION

        "This object allows the administrator to assign a unique
        operator identifier also called MPLS-TP Global_ID."
    REFERENCE
        "MPLS-TP Identifiers [TPIDS]."
    ::= { mplsIdObjects 1 }

mplsIcc OBJECT-TYPE
    SYNTAX      MplsIccId
    MAX-ACCESS  read-write
    STATUS      current
    DESCRIPTION
        "This object allows the operator or service provider to
        assign a unique MPLS-TP ITU-T Carrier Code (ICC) to a
        network."
    REFERENCE
        "MPLS-TP Identifiers [TPIDS]."
    ::= { mplsIdObjects 2 }

mplsNodeId OBJECT-TYPE
    SYNTAX      MplsNodeId
    MAX-ACCESS  read-write
    STATUS      current
    DESCRIPTION
        "This object allows the operator or service provider to
        assign a unique MPLS-TP Node_ID.

        The Node_ID is assigned within the scope of the
        Global_ID."

    REFERENCE
        "MPLS-TP Identifiers [TPIDS]."
    ::= { mplsIdObjects 3 }

-- Module compliance.

mplsIdGroups
    OBJECT IDENTIFIER ::= { mplsIdConformance 1 }
```

```
mplsIdCompliances
  OBJECT IDENTIFIER ::= { mplsIdConformance 2 }

-- Compliance requirement for fully compliant implementations.

mplsIdModuleFullCompliance MODULE-COMPLIANCE
  STATUS current
  DESCRIPTION
    "Compliance statement for agents that provide full
    support the MPLS-ID-STD-MIB module."

  MODULE -- this module

    -- The mandatory group has to be implemented by all
    -- LSRs that originate/terminate MPLS-TP paths.

    MANDATORY-GROUPS {
      mplsIdScalarGroup
    }

    ::= { mplsIdCompliances 1 }

-- Compliance requirement for read-only implementations.

mplsIdModuleReadOnlyCompliance MODULE-COMPLIANCE
  STATUS current
  DESCRIPTION
    "Compliance statement for agents that provide full
    support the MPLS-ID-STD-MIB module."

  MODULE -- this module

    -- The mandatory group has to be implemented by all
    -- LSRs that originate/terminate MPLS-TP paths.

    MANDATORY-GROUPS {
      mplsIdScalarGroup
    }

    ::= { mplsIdCompliances 2 }

-- Units of conformance.

mplsIdScalarGroup OBJECT-GROUP
  OBJECTS { mplsGlobalId,
            mplsNodeId,
            mplsIcc
```

```
}
STATUS current
DESCRIPTION
    "Scalar object needed to implement MPLS TP path."
 ::= { mplsIdGroups 1 }

-- MPLS-ID-STD-MIB module ends
END
```

12. MPLS LSR Extension MIB definitions

```
MPLS-LSR-EXT-STD-MIB DEFINITIONS ::= BEGIN

IMPORTS
    MODULE-IDENTITY, OBJECT-TYPE
        FROM SNMPv2-SMI -- [RFC2578]
    MODULE-COMPLIANCE, OBJECT-GROUP
        FROM SNMPv2-CONF -- [RFC2580]
    mplsStdMIB
        FROM MPLS-TC-STD-MIB -- [RFC3811]
    RowPointer
        FROM SNMPv2-TC -- [RFC2579]
    mplsXCIndex, mplsXCInSegmentIndex, mplsXCOutSegmentIndex,
    mplsInSegmentGroup, mplsOutSegmentGroup, mplsXCGroup,
    mplsPerfGroup, mplsLsrNotificationGroup
        FROM MPLS-LSR-STD-MIB; -- [RFC3813]

mplsLsrExtStdMIB MODULE-IDENTITY
    LAST-UPDATED
        "201106160000Z" -- June 16, 2011
    ORGANIZATION
        "Multiprotocol Label Switching (MPLS) Working Group"
    CONTACT-INFO
        "
            Venkatesan Mahalingam
            Aricent,
            India
            Email: venkatesan.mahalingam@aricent.com

            Kannan KV Sampath
            Aricent,
            India
            Email: Kannan.Sampath@aricent.com

            Sam Aldrin
            Huawei Technologies
            2330 Central Express Way,
            Santa Clara, CA 95051, USA
```

Email: aldrin.ietf@gmail.com

Thomas D. Nadeau
CA Technologies
273 Corporate Drive, Portsmouth, NH, USA
Email: thomas.nadeau@ca.com

"

DESCRIPTION

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

This MIB module contains generic object definitions for MPLS LSR in transport networks."

-- Revision history.

REVISION

"201106160000Z" -- June 16, 2011

DESCRIPTION

"MPLS LSR specific mib objects extension"

::= { mplsStdMIB xxx } -- xxx to be replaced with correct value

-- traps

mplsLsrExtNotifications OBJECT IDENTIFIER ::= { mplsLsrExtStdMIB 0 }

-- tables, scalars

mplsLsrExtObjects OBJECT IDENTIFIER ::= { mplsLsrExtStdMIB 1 }

-- conformance

mplsLsrExtConformance OBJECT IDENTIFIER ::= { mplsLsrExtStdMIB 2 }

-- MPLS LSR common objects

mplsXCExtTable OBJECT-TYPE

SYNTAX SEQUENCE OF MplsXCExtEntry

MAX-ACCESS not-accessible

STATUS current

DESCRIPTION

"This table sparse augments the mplsXCTable of MPLS-LSR-STD-MIB [RFC 3813] to provide MPLS-TP specific information about associated tunnel information"

REFERENCE

"1. Multiprotocol Label Switching (MPLS) Label Switching Router (LSR) Management Information Base (MIB), RFC 3813."

::= { mplsLsrExtObjects 1 }

mplsXCExtEntry OBJECT-TYPE

SYNTAX MplsXCExtEntry

MAX-ACCESS not-accessible

```

STATUS          current
DESCRIPTION
  "An entry in this table extends the cross connect
  information represented by an entry in
  the mplsXCTable in MPLS-LSR-STD-MIB [RFC 3813] through
  a sparse augmentation.  An entry can be created by a network
  administrator via SNMP SET commands, or in
  response to signaling protocol events."
REFERENCE
  "1. Multiprotocol Label Switching (MPLS) Label Switching
  Router (LSR) Management Information Base (MIB), RFC 3813."
INDEX { mplsXCIndex, mplsXCInSegmentIndex,
        mplsXCOutSegmentIndex }
 ::= { mplsXCExtTable 1 }

MplsXCExtEntry ::= SEQUENCE {
    mplsXCExtTunnelPointer      RowPointer
}

mplsXCExtTunnelPointer OBJECT-TYPE
    SYNTAX          RowPointer
    MAX-ACCESS      read-create
    STATUS          current
    DESCRIPTION
        "This object indicates the back pointer to the tunnel entry
        segment.  This object cannot be modified if
        mplsXCRowStatus for the corresponding entry in the
        mplsXCTable is active(1)."
    REFERENCE
        "1. Multiprotocol Label Switching (MPLS) Label Switching
        Router (LSR) Management Information Base (MIB), RFC 3813."
 ::= { mplsXCExtEntry 1 }

mplsLsrExtGroups
    OBJECT IDENTIFIER ::= { mplsLsrExtConformance 1 }
mplsLsrExtCompliances
    OBJECT IDENTIFIER ::= { mplsLsrExtConformance 2 }

-- Compliance requirement for fully compliant implementations.

mplsLsrExtModuleFullCompliance MODULE-COMPLIANCE
    STATUS current
    DESCRIPTION
        "Compliance statement for agents that provide full support
        for MPLS-LSR-EXT-STD-MIB.

        The mandatory group has to be implemented by all LSRs that
        originate, terminate, or act as transit for TE-LSPs/tunnels."

```

In addition, depending on the type of tunnels supported, other groups become mandatory as explained below."

```
MODULE MPLS-LSR-STD-MIB -- The MPLS-LSR-STD-MIB, RFC3813
```

```
MANDATORY-GROUPS {  
    mplsInSegmentGroup,  
    mplsOutSegmentGroup,  
    mplsXCGroup,  
    mplsPerfGroup,  
    mplsLsrNotificationGroup  
}
```

```
MODULE -- this module
```

```
MANDATORY-GROUPS {  
    mplsXCExtGroup  
}
```

```
OBJECT      mplsXCExtTunnelPointer
```

```
SYNTAX      RowPointer
```

```
MIN-ACCESS  read-only
```

```
DESCRIPTION
```

```
    "The only valid value for Tunnel Pointer is mplsTunnelTable  
    entry."
```

```
::= { mplsLsrExtCompliances 1 }
```

```
-- Compliance requirement for implementations that provide read-only  
-- access.
```

```
mplsLsrExtModuleReadOnlyCompliance MODULE-COMPLIANCE
```

```
STATUS current
```

```
DESCRIPTION
```

```
    "Compliance requirement for implementations that only provide  
    read-only support for MPLS-LSR-EXT-STD-MIB. Such devices can  
    then be monitored but cannot be configured using this  
    MIB module."
```

```
MODULE MPLS-LSR-STD-MIB
```

```
MANDATORY-GROUPS {  
    mplsInterfaceGroup,  
    mplsInSegmentGroup,  
    mplsOutSegmentGroup,  
}
```

```

    mplsXCGroup,
    mplsPerfGroup
}

MODULE -- this module

MANDATORY-GROUPS {
    mplsXCExtGroup
}

OBJECT      mplsXCExtTunnelPointer
SYNTAX      RowPointer
MIN-ACCESS  read-only
DESCRIPTION
    "The only valid value for Tunnel Pointer is mplsTunnelTable
    entry."

 ::= { mplsLsrExtCompliances 2 }

mplsXCExtGroup OBJECT-GROUP
OBJECTS {
    mplsXCExtTunnelPointer
}
STATUS current
DESCRIPTION
    "This object should be supported in order to access
    the tunnel entry from XC entry."
 ::= { mplsLsrExtGroups 1 }

-- MPLS-LSR-EXT-STD-MIB module ends
END

```

13. MPLS Tunnel Extension MIB definitions

```

MPLS-TE-EXT-STD-MIB DEFINITIONS ::= BEGIN

IMPORTS
    MODULE-IDENTITY, OBJECT-TYPE, Unsigned32, Gauge32,
        NOTIFICATION-TYPE
        FROM SNMPv2-SMI -- [RFC2578]
    MODULE-COMPLIANCE, OBJECT-GROUP, NOTIFICATION-GROUP
        FROM SNMPv2-CONF -- [RFC2580]
    RowStatus, StorageType
        FROM SNMPv2-TC -- [RFC2579]
    MplsLocalId, MplsGlobalId, MplsNodeId, MplsIccId
        FROM MPLS-TC-EXT-STD-MIB

```

```
mplsStdMIB, MplsTunnelIndex, MplsTunnelInstanceIndex
  FROM MPLS-TC-STD-MIB -- [RFC3811]
mplsTunnelIndex, mplsTunnelInstance, mplsTunnelIngressLSRId,
mplsTunnelEgressLSRId
  FROM MPLS-TE-STD-MIB -- [RFC3812]
;
```

mplsTeExtStdMIB MODULE-IDENTITY

LAST-UPDATED

"201106160000Z" -- June 16, 2011

ORGANIZATION

"Multiprotocol Label Switching (MPLS) Working Group"

CONTACT-INFO

"

Venkatesan Mahalingam
Aricent,
India

Email: venkatesan.mahalingam@aricent.com

Kannan KV Sampath
Aricent,

India

Email: Kannan.Sampath@aricent.com

Sam Aldrin
Huawei Technologies
2330 Central Express Way,
Santa Clara, CA 95051, USA

Email: aldrin.ietf@gmail.com

Thomas D. Nadeau
CA Technologies
273 Corporate Drive, Portsmouth, NH, USA

Email: thomas.nadeau@ca.com

"

DESCRIPTION

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

This MIB module contains generic object definitions for
MPLS Traffic Engineering in transport networks."

-- Revision history.

REVISION

"201106160000Z" -- June 16, 2011

```
DESCRIPTION
    "MPLS TE mib objects extension"

 ::= { mplsStdMIB xxx } -- xxx to be replaced with correct value

-- Top level components of this MIB module.

-- traps
mplsTeExtNotifications OBJECT IDENTIFIER ::= { mplsTeExtStdMIB 0 }
-- tables, scalars
mplsTeExtObjects          OBJECT IDENTIFIER ::= { mplsTeExtStdMIB 1 }
-- conformance
mplsTeExtConformance     OBJECT IDENTIFIER ::= { mplsTeExtStdMIB 2 }

-- Start of MPLS Transport Profile Node configuration table
mplsNodeConfigTable OBJECT-TYPE
    SYNTAX          SEQUENCE OF MplsNodeConfigEntry
    MAX-ACCESS      not-accessible
    STATUS          current
    DESCRIPTION
        "This table allows the administrator to map a node or LSR
        Identifier (IP compatible [Global_Node_ID] or ICC) with
        a local identifier.

        This table is created to reuse the existing
        mplsTunnelTable for MPLS based transport network
        tunnels also.
        Since the MPLS tunnel's Ingress/Egress LSR identifiers'
        size (Unsigned32) value is not compatible for
        MPLS-TP tunnel i.e. Global_Node_Id of size 8 bytes and
        ICC of size 6 bytes, there exists a need to map the
        Global_Node_ID or ICC with the local identifier of size
        4 bytes (Unsigned32) value in order
        to index (Ingress/Egress LSR identifier)
        the existing mplsTunnelTable."
    ::= { mplsTeExtObjects 1 }

mplsNodeConfigEntry OBJECT-TYPE
    SYNTAX          MplsNodeConfigEntry
    MAX-ACCESS      not-accessible
    STATUS          current
    DESCRIPTION
        "An entry in this table represents a mapping
        identification for the operator or service provider
        with node or LSR.

        As per [TPIDS], this mapping is
```

represented as Global_Node_ID or ICC.

Note: Each entry in this table should have a unique Global_ID and Node_ID combination."

```
INDEX { mplsNodeConfigLocalId }
 ::= { mplsNodeConfigTable 1 }
```

```
MplsNodeConfigEntry ::= SEQUENCE {
    mplsNodeConfigLocalId      MplsLocalId,
    mplsNodeConfigGlobalId    MplsGlobalId,
    mplsNodeConfigNodeId      MplsNodeId,
    mplsNodeConfigIccId       MplsIccId,
    mplsNodeConfigRowStatus   RowStatus,
    mplsNodeConfigStorageType StorageType
}
```

```
mplsNodeConfigLocalId OBJECT-TYPE
    SYNTAX      MplsLocalId
    MAX-ACCESS  not-accessible
    STATUS      current
    DESCRIPTION
        "This object allows the administrator to assign a unique
         local identifier to map Global_Node_ID or ICC."
    ::= { mplsNodeConfigEntry 1 }
```

```
mplsNodeConfigGlobalId OBJECT-TYPE
    SYNTAX      MplsGlobalId
    MAX-ACCESS  read-write
    STATUS      current
    DESCRIPTION
        "This object indicates the Global Operator Identifier.
         This object value should be zero when
         mplsNodeConfigIccId is configured with non-null value."
    REFERENCE
        "MPLS-TP Identifiers [TPIDS]."
    ::= { mplsNodeConfigEntry 2 }
```

```
mplsNodeConfigNodeId OBJECT-TYPE
    SYNTAX      MplsNodeId
    MAX-ACCESS  read-write
    STATUS      current
    DESCRIPTION
        "This object indicates the Node_ID within the operator.
         This object value should be zero when mplsNodeConfigIccId
         is configured with non-null value."
    REFERENCE
```

```
        "MPLS-TP Identifiers [TPIDS]."
 ::= { mplsNodeConfigEntry 3 }

mplsNodeConfigIccId OBJECT-TYPE
    SYNTAX      MplsIccId
    MAX-ACCESS  read-write
    STATUS      current
    DESCRIPTION
        "This object allows the operator or service provider to
        configure a unique MPLS-TP ITU-T Carrier Code (ICC)
        either for Ingress ID or Egress ID.

        This object value should be zero when
        mplsNodeConfigGlobalId and mplsNodeConfigNodeId are
        assigned with non-zero value."
    REFERENCE
        "MPLS-TP Identifiers [TPIDS]."
 ::= { mplsNodeConfigEntry 4 }

mplsNodeConfigRowStatus OBJECT-TYPE
    SYNTAX      RowStatus
    MAX-ACCESS  read-create
    STATUS      current
    DESCRIPTION
        "This object allows the administrator to create, modify,
        and/or delete a row in this table."
 ::= { mplsNodeConfigEntry 5 }

mplsNodeConfigStorageType OBJECT-TYPE
    SYNTAX      StorageType
    MAX-ACCESS  read-create
    STATUS      current
    DESCRIPTION
        "This variable indicates the storage type for this
        object.
        Conceptual rows having the value 'permanent'
        need not allow write-access to any columnar
        objects in the row."
    DEFVAL { volatile }
 ::= { mplsNodeConfigEntry 6 }

-- End of MPLS Transport Profile Node configuration table

-- Start of MPLS Transport Profile Node IP compatible mapping table

mplsNodeIpMapTable OBJECT-TYPE
    SYNTAX      SEQUENCE OF MplsNodeIpMapEntry
```

```

MAX-ACCESS      not-accessible
STATUS          current
DESCRIPTION
    "This read-only table allows the administrator to retrieve
    the local identifier for a given Global_Node_ID in an IP
    compatible operator environment.

    This table MAY be used in on-demand and/or proactive
    OAM operations to get the Ingress/Egress LSR
    identifier (Local Identifier) from Src-Global_Node_ID
    or Dst-Global_Node_ID and the Ingress and Egress LSR
    identifiers are used to retrieve the tunnel entry.

    This table returns nothing when the associated entry
    is not defined in mplsNodeConfigTable."
 ::= { mplsTeExtObjects 2 }

```

```

mplsNodeIpMapEntry OBJECT-TYPE
SYNTAX          MplsNodeIpMapEntry
MAX-ACCESS      not-accessible
STATUS          current
DESCRIPTION
    "An entry in this table represents a mapping of
    Global_Node_ID with the local identifier.

    An entry in this table is created automatically when
    the Local identifier is associated with Global_ID and
    Node_Id in the mplsNodeConfigTable.

```

```

    Note: Each entry in this table should have a unique
    Global_ID and Node_ID combination."
INDEX { mplsNodeIpMapGlobalId,
        mplsNodeIpMapNodeId
      }
 ::= { mplsNodeIpMapTable 1 }

```

```

MplsNodeIpMapEntry ::= SEQUENCE {
    mplsNodeIpMapGlobalId  MplsGlobalId,
    mplsNodeIpMapNodeId   MplsNodeId,
    mplsNodeIpMapLocalId  MplsLocalId
}

```

```

mplsNodeIpMapGlobalId OBJECT-TYPE
SYNTAX          MplsGlobalId
MAX-ACCESS      not-accessible

```

```
STATUS          current
DESCRIPTION
  "This object indicates the Global_ID."
 ::= { mplsNodeIpMapEntry 1 }

mplsNodeIpMapNodeId OBJECT-TYPE
SYNTAX          MplsNodeId
MAX-ACCESS      not-accessible
STATUS          current
DESCRIPTION
  "This object indicates the Node_ID within the
   operator."
 ::= { mplsNodeIpMapEntry 2 }

mplsNodeIpMapLocalId OBJECT-TYPE
SYNTAX          MplsLocalId
MAX-ACCESS      read-only
STATUS          current
DESCRIPTION
  "This object contains an IP compatible local identifier
   which is defined in mplsNodeConfigTable."
 ::= { mplsNodeIpMapEntry 3 }

-- End MPLS Transport Profile Node IP compatible table

-- Start of MPLS Transport Profile Node ICC based table

mplsNodeIccMapTable OBJECT-TYPE
SYNTAX          SEQUENCE OF MplsNodeIccMapEntry
MAX-ACCESS      not-accessible
STATUS          current
DESCRIPTION
  "This read-only table allows the administrator to retrieve
   the local identifier for a given ICC operator in an ICC
   operator environment.

   This table MAY be used in on-demand and/or proactive
   OAM operations to get the Ingress/Egress LSR
   identifier (Local Identifier) from Src-ICC
   or Dst-ICC and the Ingress and Egress LSR
   identifiers are used to retrieve the tunnel entry.

   This table returns nothing when the associated entry
   is not defined in mplsNodeConfigTable."
 ::= { mplsTeExtObjects 3 }

mplsNodeIccMapEntry OBJECT-TYPE
SYNTAX          MplsNodeIccMapEntry
```

```

MAX-ACCESS      not-accessible
STATUS          current
DESCRIPTION

```

"An entry in this table represents a mapping of ICC with the local identifier.

An entry in this table is created automatically when the Local identifier is associated with ICC in the mplsNodeConfigTable."

```

INDEX { mplsNodeIccMapIccId }
 ::= { mplsNodeIccMapTable 1 }

```

```

MplsNodeIccMapEntry ::= SEQUENCE {
    mplsNodeIccMapIccId      MplsIccId,
    mplsNodeIccMapLocalId   MplsLocalId
}

```

```

mplsNodeIccMapIccId OBJECT-TYPE

```

```

SYNTAX      MplsIccId
MAX-ACCESS  not-accessible
STATUS      current
DESCRIPTION

```

"This object allows the operator or service provider to configure a unique MPLS-TP ITU-T Carrier Code (ICC) either for Ingress or Egress LSR ID.

The ICC is a string of one to six characters, each character being either alphabetic (i.e. A-Z) or numeric (i.e. 0-9) characters. Alphabetic characters in the ICC should be represented with upper case letters."

```

 ::= { mplsNodeIccMapEntry 1 }

```

```

mplsNodeIccMapLocalId OBJECT-TYPE

```

```

SYNTAX      MplsLocalId
MAX-ACCESS  read-only
STATUS      current
DESCRIPTION

```

"This object contains an ICC based local identifier which is defined in mplsNodeConfigTable."

```

 ::= { mplsNodeIccMapEntry 2 }

```

```
-- End MPLS Transport Profile Node ICC based table
```

```
-- Start of MPLS Tunnel table extension
```

```
mplsTunnelExtTable OBJECT-TYPE
```

```

SYNTAX      SEQUENCE OF MplsTunnelExtEntry
MAX-ACCESS  not-accessible

```

```

STATUS          current
DESCRIPTION
    "This table represents MPLS-TP specific extensions to
    mplsTunnelTable.

    As per MPLS-TP Identifiers [TPIDS] draft, LSP_ID is

    Src-Global_Node_ID::Src-Tunnel_Num::Dst-Global_Node_ID::
    Dst-Tunnel_Num::LSP_Num for IP operator and

    Src-ICC::Src-Tunnel_Num::Dst-ICC::Dst-Tunnel_Num::LSP_Num
    for ICC operator,

    mplsTunnelTable is reused for forming the LSP_ID
    as follows,

    Source Tunnel_Num is mapped with mplsTunnelIndex,
    Source Node_ID is mapped with
    mplsTunnelIngressLSRId, Destination Node_ID is
    mapped with mplsTunnelEgressLSRId LSP_Num is mapped with
    mplsTunnelInstance.

    Source Global_Node_ID and/or ICC and Destination
    Global_Node_ID and/or ICC are maintained in the
    mplsNodeConfigTable and mplsNodeConfigLocalId is
    used to create an entry in mplsTunnelTable."
REFERENCE
    "MPLS-TP Identifiers [TPIDS]."
 ::= { mplsTeExtObjects 4 }

mplsTunnelExtEntry OBJECT-TYPE
SYNTAX          MplsTunnelExtEntry

MAX-ACCESS      not-accessible
STATUS          current
DESCRIPTION
    "An entry in this table represents MPLS-TP
    specific additional tunnel configurations."
INDEX {
    mplsTunnelIndex,
    mplsTunnelInstance,
    mplsTunnelIngressLSRId,
    mplsTunnelEgressLSRId
}
 ::= { mplsTunnelExtTable 1 }

```

```

MplsTunnelExtEntry ::= SEQUENCE {
    mplsTunnelExtDestTnlIndex  MplsTunnelIndex,
    mplsTunnelExtDestTnlLspIndex  MplsTunnelInstanceIndex
}

```

mplsTunnelExtDestTnlIndex OBJECT-TYPE
 SYNTAX MplsTunnelIndex
 MAX-ACCESS read-create
 STATUS current
 DESCRIPTION
 "This object is applicable only for the bidirectional tunnel that has the forward and reverse LSPs in the same tunnel or in the different tunnels.

This object holds the same value as that of the mplsTunnelIndex of mplsTunnelEntry if the forward and reverse LSPs are in the same tunnel. Otherwise, this object holds the value of the other direction associated LSP's mplsTunnelIndex from a different tunnel.

The values of this object and the mplsTunnelExtDestTnlLspIndex object together can be used to identify an opposite direction LSP i.e. if the mplsTunnelIndex and mplsTunnelInstance hold the value for forward LSP, this object and mplsTunnelExtDestTnlLspIndex can be used to retrieve the reverse direction LSP and vice versa.

This object and mplsTunnelExtDestTnlLspIndex values provide the first two indices of tunnel entry and the remaining indices can be derived as follows, if both the forward and reverse LSPs are present in the same tunnel, the opposite direction LSP's Ingress and Egress Identifier will be same for both the LSPs, else the Ingress and Egress Identifiers should be swapped in order to index the other direction tunnel.

The value of zero for this object is invalid."
 ::= { mplsTunnelExtEntry 1 }

mplsTunnelExtDestTnlLspIndex OBJECT-TYPE
 SYNTAX MplsTunnelInstanceIndex
 MAX-ACCESS read-create
 STATUS current
 DESCRIPTION
 "This object is applicable only for the bidirectional tunnel that has the forward and reverse LSPs in the same tunnel or in the different tunnels.

This object should contain different value if both the forward and reverse LSPs present in the same tunnel.

This object can contain same value or different values if the forward and reverse LSPs present in the different tunnels.

The value of zero for this object is valid for the configured tunnel."

```
::= { mplsTunnelExtEntry 2 }
```

```
-- End of MPLS Tunnel table extension
```

```
-- Notifications.
```

```
-- Notifications objects need to be added here.
```

```
-- End of notifications.
```

```
-- Module compliance.
```

```
mplsTeExtGroups
```

```
  OBJECT IDENTIFIER ::= { mplsTeExtConformance 1 }
```

```
mplsTeExtCompliances
```

```
  OBJECT IDENTIFIER ::= { mplsTeExtConformance 2 }
```

```
-- Compliance requirement for fully compliant implementations.
```

```
mplsTeExtModuleFullCompliance MODULE-COMPLIANCE
```

```
  STATUS current
```

```
  DESCRIPTION
```

```
    "Compliance statement for agents that provide full support the MPLS-TE-EXT-STD-MIB module."
```

```
  MODULE -- this module
```

```
-- The mandatory group has to be implemented by all  
-- LSRs that originate/terminate MPLS-TP tunnels.  
-- In addition, depending on the type of tunnels  
-- supported, other groups become mandatory as  
-- explained below.
```

```
MANDATORY-GROUPS {  
  mplsTunnelExtGroup  
}
```

```
GROUP mplsTunnelExtIpOperatorGroup
```

```
DESCRIPTION
    "This group is mandatory for devices which support
    configuration of IP based identifier tunnels."

GROUP mplsTunnelExtIccOperatorGroup
DESCRIPTION
    "This group is mandatory for devices which support
    configuration of ICC based tunnels."

 ::= { mplsTeExtCompliances 1 }

-- Compliance requirement for read-only implementations.

mplsTeExtModuleReadOnlyCompliance MODULE-COMPLIANCE
STATUS current
DESCRIPTION
    "Compliance statement for agents that provide full
    support the MPLS-TE-EXT-STD-MIB module."

MODULE -- this module

-- The mandatory group has to be implemented by all
-- LSRs that originate/terminate MPLS-TP tunnels.
-- In addition, depending on the type of tunnels
-- supported, other groups become mandatory as
-- explained below.

MANDATORY-GROUPS {
    mplsTunnelExtGroup
}

GROUP mplsTunnelExtIpOperatorGroup
DESCRIPTION
    "This group is mandatory for devices which support
    configuration of IP based identifier tunnels."

GROUP mplsTunnelExtIccOperatorGroup

DESCRIPTION
    "This group is mandatory for devices which support
    configuration of ICC based tunnels."

 ::= { mplsTeExtCompliances 2 }

-- Units of conformance.
```

```
mplsTunnelExtGroup OBJECT-GROUP
  OBJECTS {
    mplsTunnelExtDestTnlIndex,
    mplsTunnelExtDestTnlLspIndex
  }
  STATUS current
  DESCRIPTION
    "Necessary, but not sufficient, set of objects to
    implement tunnels. In addition, depending on the
    operating environment, the following groups are
    mandatory."
  ::= { mplsTeExtGroups 1 }

mplsTunnelExtIpOperatorGroup OBJECT-GROUP
  OBJECTS { mplsNodeConfigGlobalId,
    mplsNodeConfigNodeId,
    mplsNodeConfigRowStatus,
    mplsNodeConfigStorageType,
    mplsNodeIpMapLocalId
  }
  STATUS current
  DESCRIPTION
    "Object(s) needed to implement IP compatible tunnels."
  ::= { mplsTeExtGroups 2 }

mplsTunnelExtIccOperatorGroup OBJECT-GROUP
  OBJECTS { mplsNodeConfigIccId,
    mplsNodeConfigRowStatus,
    mplsNodeConfigStorageType,
    mplsNodeIccMapLocalId
  }
  STATUS current
  DESCRIPTION
    "Object(s) needed to implement ICC based tunnels."
  ::= { mplsTeExtGroups 3 }

-- MPLS-TE-EXT-STD-MIB module ends
END
```

14. Security Consideration

There is a number of management objects defined in this MIB module that has a MAX-ACCESS clause of read-write.. Such objects may be considered sensitive or vulnerable in some network environments. The support for SET operations in a non-secure environment without proper protection can have a negative effect on network operations.

Some of the readable objects in this MIB module (i.e., objects with a MAX-ACCESS other than not-accessible) may be considered sensitive or vulnerable in some network environments. It is thus important to control even GET and/or NOTIFY access to these objects and possibly to even encrypt the values of these objects when sending them over the network via SNMP. These are the tables and objects and their sensitivity/vulnerability:

SNMP versions prior to SNMPv3 did not include adequate security. Even if the network itself is secure (for example by using IPsec), even then, there is no control as to who on the secure network is allowed to access and GET/SET (read/change/create/delete) the objects in this MIB module.

It is RECOMMENDED that implementers consider the security features as provided by the SNMPv3 framework (see [RFC3410], section 8), including full supports for the SNMPv3 cryptographic mechanisms (for authentication and privacy).

Further, deployment of SNMP versions prior to SNMPv3 is NOT RECOMMENDED. Instead, it is RECOMMENDED to deploy SNMPv3 and to enable cryptographic security. It is then a customer/operator responsibility to ensure that the SNMP entity giving access to an instance of this MIB module is properly configured to give access to the objects only to those principals (users) that have legitimate rights to indeed GET or SET (change/create/delete) them.

15. IANA Considerations

To be added in a later version of this document.

16. References

16.1 Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [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, January 2001.

16.2 Informative References

[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.

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

[RFC3811] Nadeau, T., Ed., and J. Cucchiara, Ed., "Definitions of Textual Conventions (TCs) for Multiprotocol Label Switching (MPLS) Management", RFC 3811, June 2004.

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

[TPIDS] M. Bocci, et al, "MPLS-TP Identifiers", draft-ietf-mpls-tp-identifiers-03, October 25, 2010

17. Acknowledgments

To be added in a later version of this document.

18. Authors' Addresses

Sam Aldrin
Huawei Technologies
2330 Central Express Way,
Santa Clara, CA 95051, USA
Email: aldrin.ietf@gmail.com

Thomas D. Nadeau
CA Technologies
273 Corporate Drive, Portsmouth, NH, USA

Email: thomas.nadeau@ca.com

Venkatesan Mahalingam
Aricent
India
Email: venkatesan.mahalingam@aricent.com

Kannan KV Sampath
Aricent
India
Email: Kannan.Sampath@aricent.com

MPLS Working Group
Internet Draft
Intended Status: Standards Track
Expires: January 2012

S. Kini
S. Narayanan
Ericsson
July 11, 2011

MPLS Fast Re-route using extensions to LDP
draft-kini-mpls-frr-ldp-01.txt

Abstract

LDP is widely deployed in MPLS networks to signal LSPs. Since LDP establishes LSPs along IGP routed paths, its failure recovery is gated by IGP's re-convergence. Mechanisms such as IPFRR and RSVP-TE based FRR have been used to provide faster re-route for LDP LSPs. However these techniques have significant complexity and/or may not have full coverage. In this document we describe a method to perform fast re-route of LDP LSPs. The goal is to have recovery characteristics similar to the methods in [RSVP-TE-FRR] without depending on additional protocols but at the same time provide full coverage.

Status of this Memo

Distribution of this memo is unlimited.

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/lid-abstracts.html>

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

Copyright and License Notice

Copyright (c) 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	4
2. Scope	4
3. Terminology	4
4. LDP Local Repair Technique	4
4.1. Per-prefix protection	5
4.1.1 Sharing BSP LSPs using label stacking	6
4.1.2. Shortest-path LSPs as BSP LSPs	6
4.1.3. Hybrid BSP using shortest-paths	7
4.2. Per nexthop protection	7
5. Protocol extensions	8
5.1. Failure Element TLV	8
5.2. Backup Path Vector TLV	8
5.3. Tunneled FEC TLV	9
5.3. Protocol procedures	9
6. Security Considerations	10
7. IANA Considerations	10
8. References	10
8.1. Normative References	10
8.2. Informative References	10
8. Acknowledgements	10
Authors' Addresses	11

1. Introduction

LDP is a widely deployed signaling protocol in MPLS networks. It signals LSPs along IGP routed paths. In case of a failure in the network, the recovery of traffic on LDP LSPs is gated by re-convergence of IGPs. IGPs have relatively slower convergence since it is affected by factors such as link-state database flooding, re-computation etc. Approaches such as [IPFRR-LFA] can provide an alternate route that may be used by LDP. However this method does not provide full coverage. Other IPFRR methods such as [NOT-VIA] involve significant complexity. Another approach to protect LDP LSPs is to use RSVP-TE LSPs to the next-hop or next-next-hop and protect the LDP traffic by using the techniques specified in [RSVP-TE-FRR]. This has the complexity of deploying an additional protocol [RSVP-TE] in order to protect LDP LSPs.

In this document we describe a local-repair mechanism that can provide fast-reroute for LDP LSPs without requiring additional mechanisms from other protocols. This mechanism is henceforth referred to as FRR-LDP. It aims to provide traffic recovery times similar to that provided by [RSVP-TE-FRR]. This mechanism works for a link-state IGP such as [OSPF] and [ISIS].

2. Scope

This draft is applicable only when per platform label spaces are used. Per interface label spaces are out of scope.

3. Terminology

SPT: Shortest Path Tree

PLR: Point of Local Repair. The head-end LSR of a backup-SP LSP.

Backup-SP LSP (BSP LSP): An LDP LSP that provides a backup for a specific failure on the shortest path LDP LSP. The failed entity may be a link, a node or a SRLG. This LSP originates from the PLR(s).

Backup-SP Merge Point (BSP-MP): The LSR where the Backup-SP LSP is label switched to a label allocated for the shortest path LDP LSP. It need not be downstream of the failed entity.

Exclude-SPT: The shortest path tree from a PLR to a destination, when a particular failure point (link, node, SRLG) is excluded from the topology.

4. LDP Local Repair Technique

When a failure occurs in an IGP network, traffic along a shortest-path LSP that is upstream from the failure gets affected. Traffic along the shortest-path LSP that is not upstream of the failure does not get affected. A backup shortest-path LSP (BSP LSP) is setup from the PLR to an LSR that can label-switch the traffic back along the shortest-path LSP that is not upstream from the failure. Such an LSR is referred to as the BSP Merge Point or BSP-MP. LDP is used to setup the BSP LSPs. The BSP LSP becomes a single hop LSP when a Loop Free Alternate (LFA) is present. In such cases the mechanism in [IPFRR-LFA] are used. The mechanisms in this draft should be used when an LFA is not available.

The BSP LSPs to be setup in a network depends on the network topology, the failures that need to be protected and the method used by the PLR to protect the traffic. Link, node and SRLG failures can be protected by applying FRR-LDP. The different ways of applying FRR-LDP are described below. In all these methods none of the LSRs along the BSP LSP other than the PLR have to perform any special operation at the instant of failure in order to protect the traffic. The PLR pre-computes the label-operation actions to be performed on the failure trigger. This is due to FRR-LDP being a local protection mechanism.

4.1. Per-prefix protection

In this method BSP LSPs must be setup per prefix for every failure that needs to be protected. The BSP LSP can be along the shortest path to the prefix from the PLR, computed in the topology with the failure entity removed (Exclude-SPT). The BSP LSP terminates at the BSP MP which is the LSR along that path that is not upstream from the failure (at the PLR) and hence can label-switch the traffic from the BSP LSP to the shortest-path LSP of that prefix in the current network topology. All the LSRs along the BSP LSP, except the PLR, allocate a label for the BSP LSP. The PLR installs a label-swap operation on a local failure to switch the traffic from the shortest-path LSP of the protected prefix to the label allocated by the nexthop along the BSP LSP.

This method is illustrated through an example in Fig 1. Say all the links are of equal cost. Let L:X-y denote the label allocated by LSR X for prefix y for the shortest-path LSP. Let L:X-y-F denote the label allocated by LSR X for the prefix y for the failure F. These labels are for the BSP LSP. Consider a prefix e at LSR E. Say this prefix has to be protected against failure of link BE. In this case L:C-e-BE is a label allocated by C for the BSP LSP for the prefix p for the failure of link BE. Though D is the LSR that will merge the traffic back to the shortest-path LSP, it does not need to advertise a separate label since the shortest-path LSP label L:D-e can be re-

used. A label-swap action of L:C-e-BE to L:D-e is installed by LSR C in its ILM. Similarly, a label L:C-e-DE is allocated by LSR C. In this topology two extra labels are adequate to protect against any failure for a prefix terminating at E.

The label on the packet as it goes from C to E along the path C-B-E before the failure would be L:B-e -> L:E-e. When PLR B receives a trigger of link BE failure it installs a swap operation from L:B-e to L:C-e-BE. Thus the labels on the packet as it goes from C to E would be L:B-e -> L:C-e-BE -> L:D-e -> L:E-e.

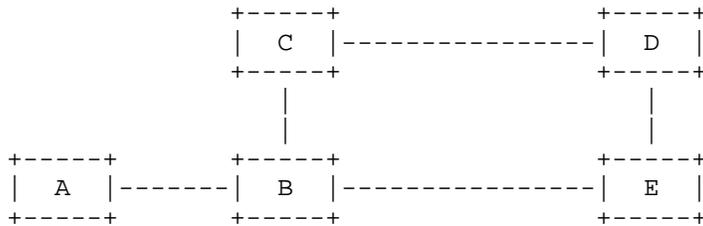


Figure 1 Example topology

4.1.1 Sharing BSP LSPs using label stacking

The number of BSP LSPs can be greatly reduced by using label stacking. When the path of BSP LSPs is the same for multiple prefixes and/or failures, the same BSP LSP can be used using label stacking. The BSP MP advertises a label for the prefix to the PLR. The protocol mechanisms to advertise this are described later. In cases where the BSP MP has multiple equal-cost paths to the destination and some of them can go through the failure point, this label must be different than the shortest-path LSP label that it has allocated for that prefix. The label-swap operations for that label have to be set such that the paths taken do not go through the failure point. During a protection switch the PLR does a label swap of the protected LSP to the label advertised by the BSP MP and then pushes the label of the shortest-path LSP to the BSP MP.

As a failure trigger action the PLR swaps the protected LSPs label with that advertised by the BSP MP for that prefix. It then pushes the BSP LSP label and forwards the packet.

4.1.2. Shortest-path LSPs as BSP LSPs

A shortest-path LSP in the current network topology from the PLR can also be used as a BSP LSP by applying label stacking at the PLR. This can be done only if such a shortest-path LSP exists and it does not

fate-share with the failure that is being protected against. The BSP MP advertises a label for that prefix to the PLR. The protocol mechanisms to advertise this are described later. In cases where the BSP MP has multiple equal-cost paths to the destination and some of them can go through the failure point, this label must be different than the shortest-path LSP label that it has allocated for that prefix. The label-swap operations for that label have to be set such that the paths taken do not go through the failure point. During a protection switch the PLR does a label swap of the protected LSP to the label advertised by the BSP MP and then pushes the label of the shortest-path LSP to the BSP MP. Such a shortest-path LSP may be used to as a BSP LSP for many prefixes by first swapping with label specific to the prefix before pushing the label of the shortest-path LSP.

In the example in Fig 1, LSR B uses the shortest-path LSP to D as the BSP LSP for protecting the LSP to prefix e from failure of the link BE. The PLR B pre-installs a failure action for link BE that it should first swap the label L:B-e to L:D-e and then push the label for the shortest-path LSP to D. In this example, there are no additional labels allocated for protection than the shortest-path LSP labels. Also multiple prefixes terminating at E can be protected using the same BSP LSP.

4.1.3. Hybrid BSP using shortest-paths

When the BSP LSP cannot use a shortest-path LSP then all the LSRs along the BSP LSP except the PLR must allocate labels. It is possible to reduce the number of labels required for the BSP LSP by encapsulating into shortest-path LSPs along the path of the BSP LSP. This further reduces the additional info needed for FRR-LDP. Details of this will be provided in future versions of the doc.

4.2. Per nexthop protection

In this method only the next-hop is protected using FRR-LDP. In this case label stacking must be used at the PLR before sending the packet to the next downstream LSR (for link protection) or the next-to-next downstream LSR (for node protection). For SRLG protection, it may be necessary to learn a label allocated by a node further downstream. Protocol extensions similar to record label would need to be defined for this.

The BSP LSP can be along the shortest path from the PLR to the next downstream LSR(or next-next-hop), computed in the topology with the failure entity removed. The BSP MP is the LSR that label-switches traffic to a shortest-path LSP to that nexthop in the current network topology.

The advantage of this method compared to per-prefix protection is that lesser number of BSP LSPs are needed. This is due to all LSPs that go through that nexthop being protected using a single BSP LSP. However the packet may take a less optimal path till the shortest-path re-computation is done.

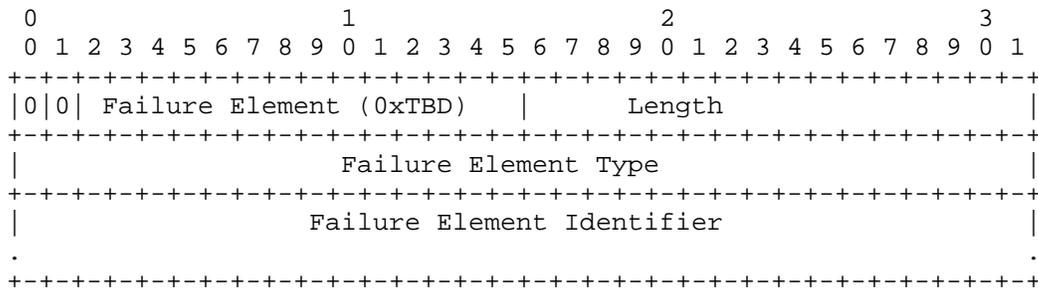
If a shortest-path LSP from the PLR to the BSP MP exists that can be used as the BSP-LSP, an additional level of label-stacking must be used. This can further reduce the number of additional labels required for BSP LSPs.

5. Protocol extensions

The BSP LSP is signaled using extensions to LDP. When a BSP LSP is used to protect multiple prefixes (by label stacking) the label mapping of each of the protected prefixes needs to be advertised to the PLR. A targeted LDP session can be used for this. This can have scalability issues and hence an option to advertise it without targeted LDP is defined.

5.1. Failure Element TLV

A Failure Element TLV identifies the failure that the BSP LSP is protecting against. It identifies that this message is for a BSP LSP.



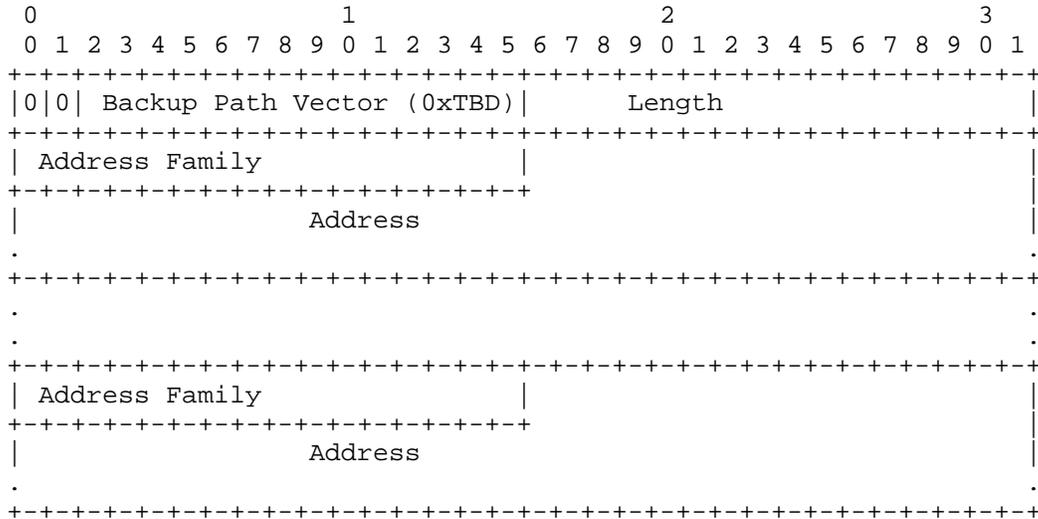
Failure Element Type Can be one of either a link (v4/v6), node (v4/v6) or a SRLG

Failure Element Identifier A link is identified by an IP address of one of its ends. A node is identified by its loopback IP address. The SRLG is as defined in RFC 4202.

5.2. Backup Path Vector TLV

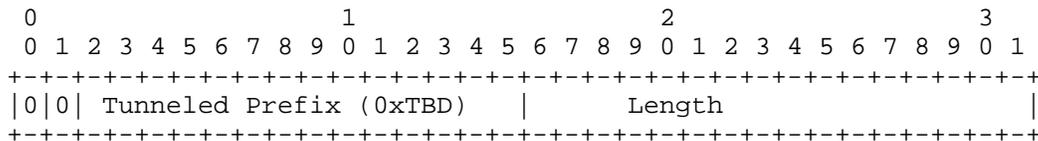
The Backup Path Vector TLV indicates the path taken by this BSP LSP from the BSP MP to the PLR. It consists of loopback addresses of each LSR along the path. The first address would be that of the MP and the

last address would be the PLR address.



5.3. Tunneled FEC TLV

The Tunneled FEC TLV indicates to the PLR the label advertised by the MP for the FEC for re-routing traffic. This label should be used to tunnel through the BSP LSP. The intermediate nodes do not install any data-plane state for a tunneled FEC.



5.3. Protocol procedures

An LSR computes the failures and prefixes for which it can act as a BSP MP and advertises a label mapping for the BSP LSP by including the Failure Element TLV and the Backup Path Vector TLV. The intermediate LSRs of a BSP LSP allocate labels for the BSP LSP and also advertise it upstream using the Backup Path Vector TLV. If label stacking is used BSP MP also advertises label mappings for the tunneled prefixes by including the Tunneled Prefix TLV in addition to the Failure Element TLV and the Backup Path Vector TLV. The intermediate LSRs do not allocate labels for this since the label is tunneled in a BSP LSP. They forward the label mapping using the Backup Path Vector TLV. The PLR installs actions for a failure trigger using the labels.

6. Security Considerations

This document does not introduce any additional security considerations beyond those in [LDP].

7. IANA Considerations

New TLV types are needed for Failure Element TLV, Backup Path Vector TLV and Tunneled Prefix TLV.

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.
- [LDP] Andersson, L., et al, "LDP Specification", RFC 5036, October 2007.
- [OSPF] Moy, J., "OSPF Version 2", STD 54, RFC 2328, April 1998.
- [ISIS] International Organization for Standardization, "Intermediate system to intermediate system intra-domain-routing routine information exchange protocol for use in conjunction with the protocol for providing the connectionless-mode Network Service (ISO 8473)", ISO Standard 10589, 1992.
- [RSVP-TE] Awduche, D., et al, "RSVP-TE: Extensions to RSVP for LSP Tunnels", RFC 3209, December 2001.
- [RSVP-TE-FRR] Pan, P., et al, "Fast Reroute Extensions to RSVP-TE for LSP Tunnels", RFC 4090, May 2005.
- [IPFRR-LFA] Atlas, A., et al, "Basic Specification for IP Fast Reroute: Loop-Free Alternates", RFC 5286, September 2008.

8.2. Informative References

- [NOT-VIA] Shand, M., et al, "IP Fast Reroute Using Not-via Addresses", draft-ietf-rtgwg-ipfrr-notvia-addresses-07 (Work in progress), April 2011.

8. Acknowledgements

The authors would like to thank Joel Halpern for their review and

useful comments.

Authors' Addresses

Sriganesh Kini
EMail: sriganesh.kini@ericsson.com

Srikanth Narayanan
EMail: srikanth.narayanan@ericsson.com

MPLS Working Group
Internet Draft

A. D'Alessandro
Telecom Italia
M.Paul
Deutsche Telekom
S. Ueno
NTT Communications
Y.Koike
NTT

Intended status: Informational

Expires: January 13, 2012

July 11, 2011

Temporal and hitless path segment monitoring
draft-koike-mpls-tp-temporal-hitless-psm-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/lid-abstracts.txt>

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

This Internet-Draft will expire on 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.

Abstract

The MPLS transport profile (MPLS-TP) is being standardized to enable carrier-grade packet transport and complement converged packet network deployments. One of the most attractive features of MPLS-TP are OAM functions, which enable network operators or service providers to provide various maintenance characteristics, such as fault location, survivability, performance monitoring, and preliminary or in-service measurements.

One of the most important mechanisms which are common for transport network operation is fault location. A segment monitoring function of a transport path is effective in terms of extension of the maintenance work and indispensable particularly when the OAM function is effective only between end points. However, the current approach defined for MPLS-TP for the segment monitoring (SPME) has some fatal drawbacks.

This document elaborates on the problem statement for the path segment monitoring function. Moreover, this document requests to add network objectives to solve or improve the issues and proposes to consider a new improved method of segment monitoring.

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

Table of Contents

1. Introduction	4
2. Conventions used in this document.....	4
2.1. Terminology	5
2.2. Definitions	5
3. Network objectives for monitoring.....	5
4. Problem statement	5
5. OAM functions for segment monitoring	8

- 6. Further consideration of requirements for enhanced segment monitoring 10
- 6.1. Necessity of on-demand single-layer monitoring10
- 6.2. Necessity of on-demand monitoring independent from proactive monitoring 11
- 6.3. On-demand diagnostic procedures..... 11
- 7. Conclusion 12
- 8. Security Considerations..... 13
- 9. IANA Considerations 13
- 10. References 13
- 10.1. Normative References..... 13
- 10.2. Informative References..... 14
- 11. Acknowledgments 14

1. Introduction

A packet transport network will enable carriers or service providers to use network resources efficiently, reduce operational complexity and provide carrier-grade network operation. Appropriate maintenance functions, supporting fault location, survivability, performance monitoring and preliminary or in-service measurements, are essential to ensure quality and reliability of a network. They are essential in transport networks and have evolved along with TDM, ATM, SDH and OTN.

Unlike in SDH or OTN networks, where OAM is an inherent part of every frame and frames are also transmitted in idle mode, it is not possible to constantly monitor the status of individual connections in packet networks. Packet-based OAM functions are flexible and selectively configurable according to operators' needs.

According to the MPLS-TP OAM requirements [1], mechanisms MUST be available for alerting a service provider of a fault or defect affecting the service(s) provided. In addition, to ensure that faults or degradations can be localized, operators need a method to analyze or investigate the problem. From the fault localization perspective, end-to-end monitoring is insufficient. Using end-to-end OAM monitoring, when one problem occurs in an MPLS-TP network, the operator can detect the fault, but is not able to localize it.

Thus, a specific segment monitoring function for detailed analysis, by focusing on and selecting a specific portion of a transport path, is indispensable to promptly and accurately localize the fault point.

For MPLS-TP, a path segment monitoring function has been defined to perform this task. However, as noted in the MPLS-TP OAM Framework[5], the current method for segment monitoring function of a transport path has implications that hinder the usage in an operator network.

This document elaborates on the problem statement for the path segment monitoring function. Moreover, this document requests to add network objectives to solve or improve the issues and proposes to reconsider the new improved method of the segment monitoring.

2. Conventions used in this document

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

2.1. Terminology

LSP Label Switched Path
OTN Optical Transport Network
PST Path Segment Tunnel
TCM Tandem connection monitoring
SDH Synchronous Digital Hierarchy
SPME Sub-path Maintenance Element

2.2. Definitions

None

3. Network objectives for monitoring

There are two indispensable network objectives in the current on-going MPLS-TP standard as described in section 3.8 of [5].

(1) The monitoring and maintenance of current transport paths has to be conducted in service without traffic disruption.

(2) Segment monitoring must not modify the forwarding of the segment portion of the transport path.

It was agreed in ITU-T SG15 that Network objective (1) is mandatory and that regarding Network objective (2) the monitoring shall be hitless and not change the forwarding behavior.

4. Problem statement

To monitor, protect, or manage a portion of a transport path, such as LSP in MPLS-TP networks, the Sub-Path Maintenance Element (SPME) is defined in [2]. The SPME is defined between the edges of the portion of the LSP that needs to be monitored, protected, or managed. This SPME is created by stacking the shim header (MPLS header)[3] and is defined as the segment where the header is stacked. 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 by setting the TTL value of the label stack entry (LSE) and interface identifier value at the corresponding hierarchical LSP level in a per-node model.

This method has the following general issues, which are fatal in terms of cost and operation.

(P-1) Increasing the overhead by the stacking of shim header(s)

(P-2) Increasing the address management complexity, as new MEPs and MIPs need to be configured for the SPME in the old MEG

Problem (P-1) leads to decreased efficiency as bandwidth is wasted. As the size of monitored segments increases, the size of the label stack grows. Moreover, if operator wants to monitor the portion of a transport path without service disruption, one or more SPMEs have to be set in advance until the end of life of a transport path, which is not temporal or on-demand. Consuming additional bandwidth permanently for the monitoring purpose should be avoided to maximize the available bandwidth.

Problem (P-2) is related to an identifier issue, of which the discussion is still continuing and not so critical at the moment, but can be quite inefficient if the policy of allocating the identifier is different in each layer. Moreover, from the perspective of operation, increasing the managed addresses and the managed layer is not desirable in terms of simplified operation featured by current transport networks. Reducing the managed identifier and managed layer should be the fundamental direction in designing the architecture.

The most familiar example for SPME in transport networks is tandem connection monitoring (TCM), which can for example be used for a carrier's carrier solution, as shown in Fig. 17 of the framework document[2]. However, in this case, the SPMEs have to be pre-configured. If this solution is applied to specific segment monitoring within one operator domain, all the necessary specific segments have to be pre-configured. This setting increases the managed objects as well as the necessary bandwidth, shown as Problem (P-1) and (P-2).

To avoid these issues, the temporal setting of the SPME(s) only when necessary seems reasonable and efficient for monitoring in MPLS-TP transport network operation. Unfortunately, the temporal settings of SPMEs also cause the following problems due to label stacking, which are fatal in terms of intrinsic monitoring and service disruption.

(P'-1) Changing the condition of the original transport path by changing the length of the MPLS frame (delay measurement and loss measurement can be sensitive)

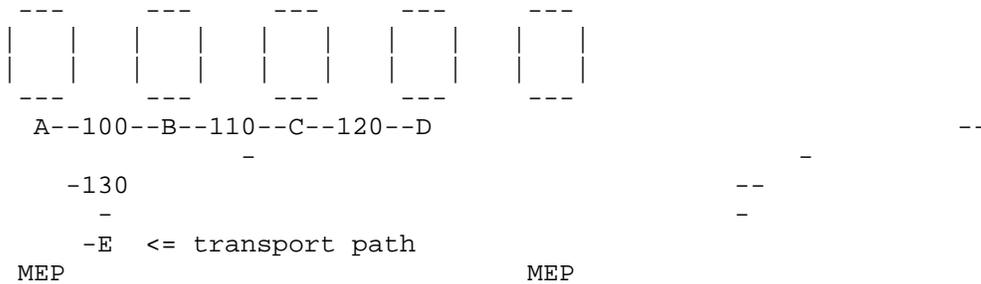
(P'-2) Disrupting client traffic over a transport path, if the SPME is temporally configured.

Problem (P'-1) is a fatal problem in terms of intrinsic monitoring. The monitoring function checks the status without changing any conditions of the targeted monitored segment or the transport path. If the conditions of the transport path change, the measured value or observed data will also change. This can make the monitoring meaningless because the result of the monitoring would no longer reflect the reality of the connection where the original fault or degradation occurred.

In addition, changing the settings of the original shim header should not be allowed because those changes correspond to creating a new portion of the original transport path, which is completely different from the original circumstances.

Figure 1 shows an example of SPME setting. In the figure, X means the one label expected on the tail end node D of the original transport path. "210" and "220" are label allocated for SPME. The label values of the original path are modified as well as the values of stacked label. This is not the monitoring of the original transport path but the monitoring of a different path.

(Before SPME settings)



(After SPME settings)

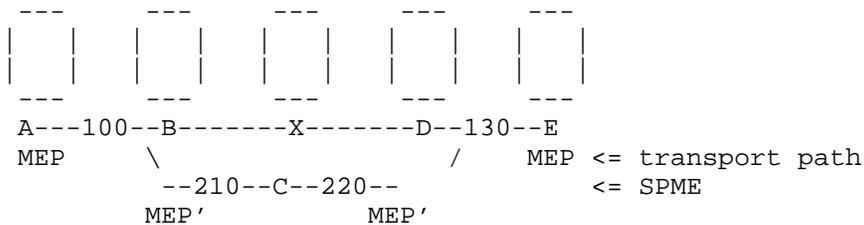


Figure 1 : An Example of a SPME setting

Problem (P'-2) was not fully discussed, although the make-before-break procedure in the survivability document [4] seemingly supports the hitless configuration for monitoring according to the framework

document [2]. The reality is the hitless configuration of SPME is impossible without affecting the circumstances of the targeted transport path, because the make-before-break procedure is premised on the change of the label value. This means changing one of the settings in MPLS shim header and should not be allowed as explained in (P'-1) above.

Moreover, this might not be effective under the static model without a control plane because the make-before-break is a restoration application based on the control plane. The removal of SPME whose segment is monitored could have the same impact (disruption of client traffics) as creation of an SPME on the same LSP.

The other potential risks are also envisaged. Setting up a temporal SPME will result in the LSRs within the monitoring segment only looking at the added (stacked) labels and not at the labels of the original LSP. This means that problems stemming from incorrect (or unexpected) treatment of labels of the original LSP by the nodes within the monitored segment may disappear when setting up SPME. This might include hardware problems during label look-up, mis-configuration etc. Therefore operators have to pay extra attention to correctly setting and checking the label values of the original LSP in the configuration. Of course, the inversion of this situation is also possible, .e.g., incorrect or unexpected treatment of SPME labels can result in false detection of a fault where none of the problem originally existed.

The way MPLS works requires use of SPMEs to be designed in advance, hence the utility of SPMEs is basically limited.

To summarize, the problem statement is that the current sub-path maintenance based on a hierarchical LSP (SPME) is problematic in terms of increasing bandwidth by label stacking and managing objects by layer stacking and address management. A temporal configuration of SPME is one of the possible approaches for minimizing the impact of these issues. However, the current method is unfavorable because the temporal configuration for monitoring can change the condition of the original monitored transport path and disrupt the in-service customer traffic. From the perspective of monitoring in transport network operation, a solution avoiding those issues or minimizing their impact is required.

5. OAM functions using segment monitoring

OAM functions in which segment monitoring is required are basically limited to on-demand monitoring which are defined in OAM framework document [5], because those segment monitoring functions are used to

locate the fault/degraded point or to diagnose the status for detailed analyses, especially when a problem occurred.

Packet loss and packet delay measurements are OAM functions in which hitless and temporal segment monitoring are strongly required because these functions are supported only between end points of a transport path. If a fault or defect occurs, there is no way to locate the defect or degradation point without using the segment monitoring function. If an operator cannot locate or narrow the cause of the fault, it is quite difficult to take prompt action to solve the problem. Therefore, temporal and hitless monitoring for packet loss and packet delay measurements are indispensable for transport network operation.

Regarding other on-demand monitoring functions, segment monitoring is desirable, but not as urgent as for packet loss and packet delay measurements.

Regarding out-of-service on-demand monitoring functions, such as diagnostic tests, there seems no need for hitless settings. However, specific segment monitoring should be applied to the OAM function of diagnostic test. See section 6.3.

Note:

The reason only on-demand OAM functions are discussed at this point is because the characteristic of "on-demand" is generally temporal for maintenance operation. Thus, operations should not be based on pre-configuration. Pre-design and pre-configuration of PST/TCM (label stacking) for all the possible patterns for the on-demand (temporal) usage are not reasonable, because these tasks will increase the operator's burden, although pre-configuration of PST for pro-active usage may be accepted considering the agreement thus far. Therefore, the solution for temporal and hitless segment monitoring does not need to be limited to label stacking mechanisms, such as PST/TCM(label stacking), which can cause the issues (P-1) and (P-2) described in Section 4.

The solution for temporal and hitless segment monitoring has to cover both per-node model and per-interface model which are specified in [5].

6. Further consideration of requirements for enhanced segment monitoring

An existing segment monitoring function relates to SPME that instantiates a hierarchical transport path (introducing MPLS label stacking) through which OAM packets can be sent. SPME construct monitoring function is particularly important mainly for protecting bundles of transport paths and carriers' carrier solutions. From this perspective, monitoring function related to can be considered ''proactive multi-layer monitoring,'' which has already been determined by consensus to be mandatory in MPLS-TP.

6.1. Necessity of on-demand single-layer monitoring

In contrast to the ''proactive multi-layer monitoring'' so called ''SPME'', the new segment monitoring function is supposed to be applied mainly for diagnostic purpose on-demand. We can differentiate this monitoring from the proactive segment monitoring as on-demand (multi-layer) monitoring. The most serious problem at the moment is that there is no way to localize the degradation point on a path without changing the conditions of the original path. Therefore, as a first step, single layer segment monitoring not affecting the monitored path is required for a new on-demand and hitless segment monitoring function.

A combination of multi-layer and simultaneous monitoring is the most powerful tool for accurately diagnosing the performance of a transport path. However, considering the balance of estimated implementation difficulties and the substantial benefits to operators, a strict monitoring function such as in a test environment in a laboratory does not seem to be necessary in the field. To summarize, on-demand and in-service (hitless) single-layer segment monitoring is required, but the need for on-demand and in-service multi-layer segment monitoring is not as urgent at the moment. Figure 2 shows an example of a multi-layer on-demand segment monitoring.

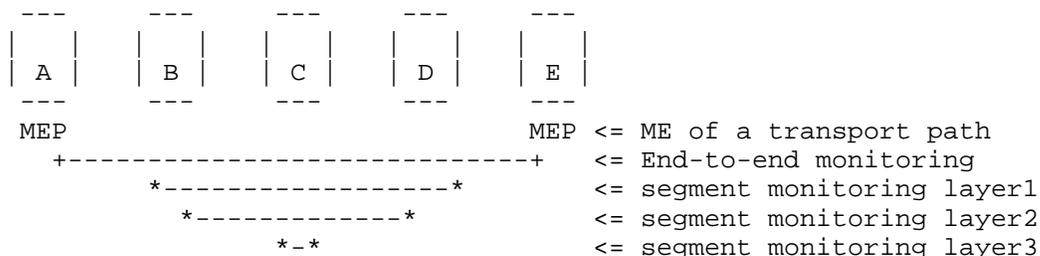


Figure 2 : An Example of a multi-layer on-demand segment monitoring

6.2. Necessity of on-demand monitoring independent from proactive monitoring

As multi-layer simultaneous monitoring only in layers for on-demand monitoring was already discussed in section 6.1, we consider the necessity of simultaneous proactive and on-demand monitoring. Normally, on-demand segment monitoring is configured in a segment of a maintenance entity of a transport path. In this environment, on-demand single-layer monitoring should be done irrespective of the status of pro-active monitoring of the targeted end-to-end transport path.

If operators have to disable the pro-active monitoring during "on-demand and in-service" segment monitoring, the network operation system might miss any performance degradation of user traffic. This kind of inconvenience should be avoided in the network operations. From this perspective, the ability for on-demand single layer segment monitoring is required without changing or interfering the proactive monitoring of the original end-to-end transport path.

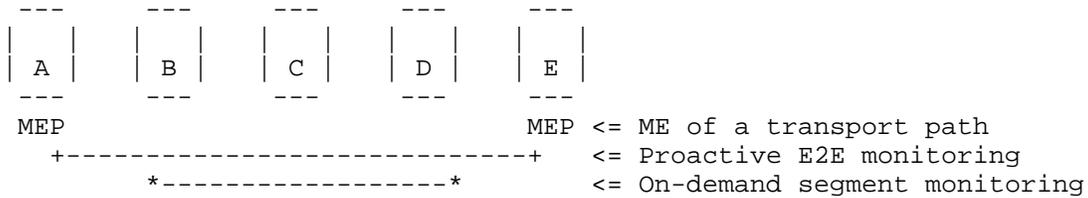


Figure 3 : Relation between proactive end-to-end monitoring and on-demand segment monitoring

6.3. On-demand diagnostic procedures

The main objective of on-demand segment monitoring is to diagnose the fault points. One possible diagnostic procedure is to fix one end point of a segment at the MEP of a transport path and change progressively the length of the segment in order. This example is shown in Fig. 4. This approach is considered as a common and realistic diagnostic procedure. In this case, one end point of a segment can be anchored at MEP at any time.

Other scenarios are also considered, one shown in Fig. 5. In this case, the operators want to diagnose a transport path from a transit node that is located at the middle, because the end nodes(A and E) are located at customer sites and consist of cost effective small box in which a subset of OAM functions are supported. In this case, if

one end point and an originator of the diagnostic packet are limited to the position of MEP, on-demand segment monitoring will be ineffective because all the segments cannot be diagnosed (For example, segment monitoring 3 in Fig.5 is not available and it is not possible to localize the fault point).

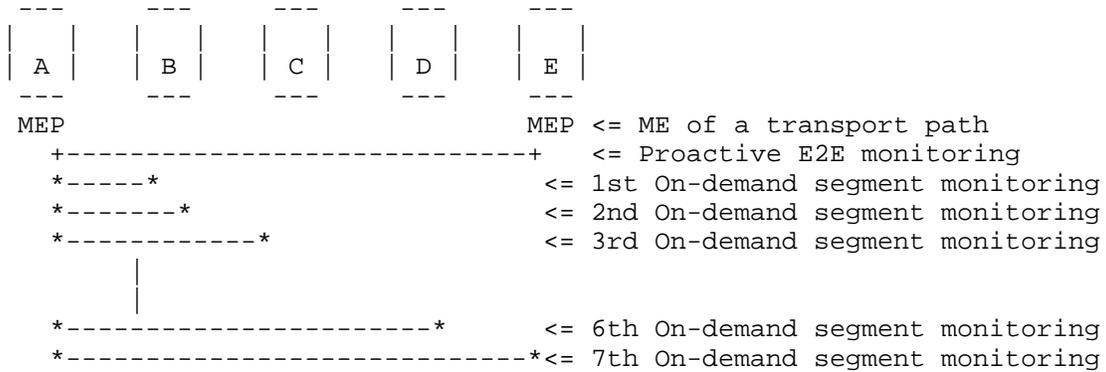


Figure 4 : One possible procedure to localize a fault point by sequential on-demand segment monitoring

Accordingly, on-demand monitoring of arbitrary segments is mandatory in the case in Fig. 5. As a result, on-demand and in-service segment monitoring should be set in an arbitrary segment of a transport path and diagnostic packets should be inserted from at least any of intermediate maintenance points of the original ME.

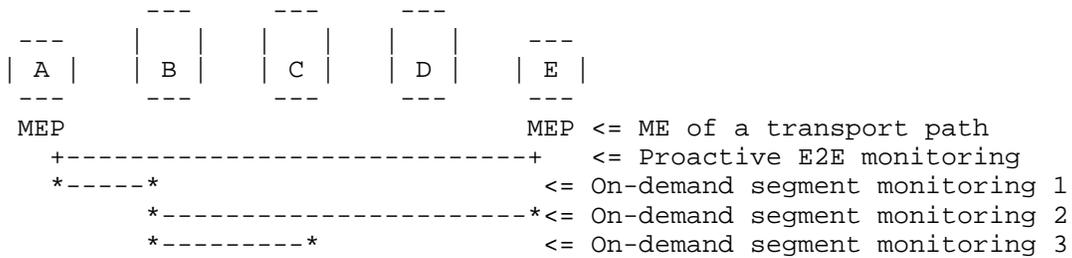


Figure 5 : Example where on-demand monitoring has to be configured in arbitrary segments

7. Conclusion

It is requested that the two network objectives mentioned in Section 3 are met by MPLS-TP OAM solutions which has been already described in section 3.8 of [5]. The enhancements should minimize the issues

described in Section 4,, i.e., P-1, P-2, P'-1 and P'-2, to meet those two network objectives. In addition, the following requirements should be considered for an enhanced temporal and hitless path segment monitoring function.

- An on-demand and in-service ''single-layer'' segment monitoring is proposed. Multi-layer segment monitoring is optional.

- ''On-demand and in-service'' single layer segment should be done without changing or interfering any condition of pro-active monitoring of an original ME of a transport path.

- On-demand and in-service segment monitoring should be able to be set in an arbitrary segment of a transport path.

8. Security Considerations

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

9. IANA Considerations

There are no IANA actions required by this draft.

10. References

10.1. Normative References

- [1] Vigoureux, M., Betts, M., Ward, D., "Requirements for OAM in MPLS Transport Networks", RFC5860, May 2010
- [2] Bocci, M., et al., "A Framework for MPLS in Transport Networks", RFC5921, July 2010
- [3] Rosen, E., et al., "MPLS Label Stack Encoding", RFC 3032, January 2001
- [4] Sprecher, N., Farrel, A. , ''Multiprotocol Label Switching Transport Profile Survivability Framework'', draft-ietf-mpls-tp-survive-fwk-06.txt(work in progress), June 2010
- [5] Busi, I., Dave, A. , "Operations, Administration and Maintenance Framework for MPLS-based Transport Networks ", draft-ietf-mpls-tp-oam-framework-11.txt(work in progress), February 2011

10.2. Informative References

None

11. Acknowledgments

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

The authors would also like to thank Alexander Vainshtein, Dave Allan, Fei Zhang, Huub van Helvoort, Italo Busi, Maarten Vissers, Malcolm Betts and Nurit Sprecher for their comments and enhancements to the text.

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

Authors' Addresses

Alessandro D'Alessandro
Telecom Italia
Email: alessandro.dalessandro@telecomitalia.it

Manuel Paul
Deutsche Telekom
Email: Manuel.Paul@telekom.de

Satoshi Ueno
NTT Communications
Email: satoshi.ueno@ntt.com

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

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: January 2, 2012

A. Lo
K. Patel
V. Lim
Cisco Systems
July 1, 2011

Incremental Label Announcement Extensions for LDP
draft-ldp-ila-extension-00.txt

Abstract

The current LDP Graceful Restart (GR) mechanism specified in RFC3478 requires a complete re-advertisement of the LDP label binding information across a session restart, even though complete label binding information might be preserved.

In this document we specify extensions to LDP graceful restart in order to support avoiding unnecessary transmission of the label binding information preserved across a session restart, thus accelerating the router re-convergence. More specifically, we describe a version number based batching mechanism for keeping track of the label binding information across a session restart.

The new 1) LDP ILA capability TLV, 2) LDP VERSION ID TLV and 3) LDP ILA Version message type, is introduced for checkpointing the label binding version maintained for a neighbor. We also specify procedures for handling label binding table version update across a session restart.

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 2, 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.

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

Table of Contents

1.	Introduction	4
1.1.	Requirements Language	4
2.	Incremental Label Announcement Capability TLV	4
3.	Incremental Label Announcement FEC Version TLV	5
4.	Incremental Label Announcement Version Message	6
5.	Procedures	7
5.1.	Session Initialization	7
5.2.	Label Mapping Sender/Receiver (LMS/LMR)	7
5.3.	ILA Version ID=0 Handling	9
5.4.	ILA Version ID Assignment	9
5.5.	ILA Version ID wraparound	9
6.	Acknowledgements	9
7.	IANA Considerations	10
8.	Security Considerations	10
9.	Normative References	10
	Authors' Addresses	10

1. Introduction

This document defines a new LDP Incremental Label Announcement extension for LDP Graceful restart. This mechanism avoids unnecessary transmission of the label binding information during session restarts and thus improve the overall router convergence.

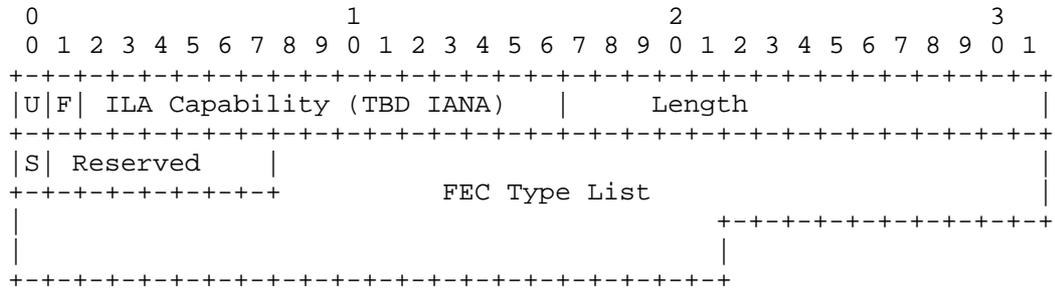
1.1. Requirements Language

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

2. Incremental Label Announcement Capability TLV

The LDP Incremental Label Announcement (ILA) Capability TLV is used by an LDP speaker to list the FEC types that support the ILA capability. This TLV MUST be announced in the LDP initialization message along with the LDP FT Session TLV. An implementation that support LDP ILA MUST implement the procedures for Capability Parameters in LDP Initialization Messages.

The format of a "Incremental Label Announcement Capability" TLV is as follows:



U-bit: The Unknown TLV bit should be set to the value 1.

F-bit: The forward unknown TLV bit should be set to the value 0.

ILA Capability: The ILA Capability code is assigned by IANA (TBD).

Length:

The length indicates the number of octets for S/Reserved byte and the bytes found in the FEC Type List Data.

S-bit:

The State Bit MUST always be set to 1. It indicates whether the sender is advertising or withdrawing the capability corresponding to the TLV code point.

The State Bit value is used as follows:

- 1 - The TLV is advertising the capability specified by the TLV code point.
- 0 - The TLV is withdrawing the capability specified by the TLV code point.

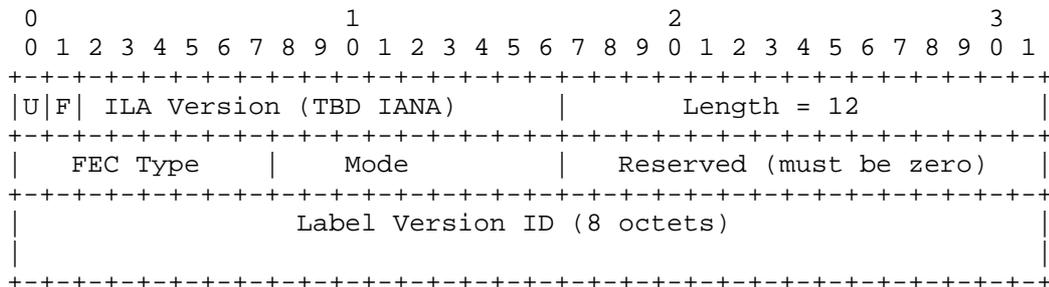
FEC Type List:

The FEC type list indicates the sender supported ILA FEC types. Each Octet of FEC Type List Data corresponds to a FEC type defined in the LDP Forwarding Equivalence Class (FEC) Type Namespace.

3. Incremental Label Announcement FEC Version TLV

The ILA Version TLV is defined for controlling/versioning label mapping advertisements/withdraw messages for a given FEC type. This TLV is used by the receiver of the label advertisements/withdraw message to request which versions of Label bindings the LDP speaker should announce from. Furthermore, it is also used by the LDP speaker to verify that the labels advertisements for a given FEC type do fall within the specified version id. The LDP speaker uses this information in generating incremental announcements.

The "ILA Version" TLV has the following format:



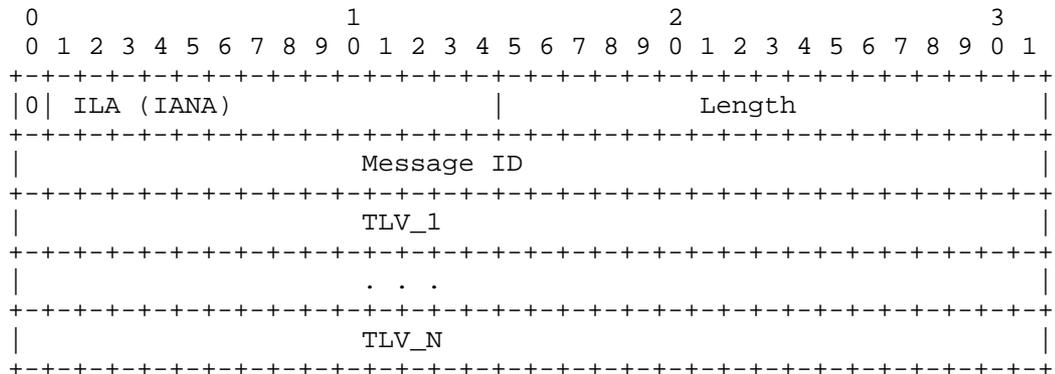
- U-bit:
The Unknown TLV bit MUST be set to the value 0.
- F-bit:
The forward unknown TLV bit MUST be set to the value 0.
- ILA Version:
The ILA Version TLV type is assigned by IANA (TBD).
- Length:
The length field is always set to 12.
- FEC type:
Identifies the FEC type for which the ILA version message applies.
- Mode:
0 - Request Mode: Request label bindings starting from specified version.
1 - Assign Mode: Assign the specified version ID to the bindings that follow.
- Label Version ID:
A 64 bit version number.

4. Incremental Label Announcement Version Message

The new Incremental Label Announcement Version message is used by the speaker to send 1 or more ILA Version TLVs. This message contains one or more per FEC type TLVs used to request the peer to start sending labels with a given version number and to inform the peer the current version ID assigned to the bindings that follow. If there are multiple TLVs of a specific FEC type, at most there MUST be one Version-id "request", and Version-ID "assign" TLV for a given FEC type.

An LDP speaker MUST not send this message unless the LDP peer has previously announced its ability to support the ILA capabilities. Only the LDP FEC types found in the ILA Capability FEC Type List should appear in the the ILA version TLVs.

The ILA Version message is defined as following:



where TLV_1 through TLV_N are currently used for ILA Version-ID TLVs.

5. Procedures

5.1. Session Initialization

An LDP speaker that is capable and willing to support the ILA procedures for a given FEC type advertises this ability through the Incremental Label Announcement Capability TLV in the LDP session initialization message. The ILA Capability TLV MUST only be included in the LDP initialization message when if LDP initialization message contains a FT session TLV indicating its ability to support LDP Graceful Restart. The sender of the ILA Capability TLV MUST include all the FEC types for which it intends to support ILA procedures for. The set of FEC types that is found in both the sent ILA Capability TLV and the received ILA Capability TLV represents the FEC types for which both LDP endpoints will follow ILA procedures when advertising/withdrawing label bindings.

The FEC type list may potentially change across a LDP restart. When this happens, the bindings for FEC types previously supporting ILA that changed disappeared for the new LDP session need to be purged.

5.2. Label Mapping Sender/Receiver (LMS/LMR)

During label mapping advertisement/withdraw, an LDP endpoint plays the role of the label mapping sender (LMS) or label mapping receiver (LMR).

For FEC types which the ILA procedures apply, the LMS will need to maintain a local label binding table and associate a ILA VERSION ID

with each binding. Each time a local label binding changes (such that a re-advertisement or withdraw needs to be sent), the VERSION ID for the binding must be updated such that the value is greater than or equal to the current VERSION_ID being used to advertise/withdraw label mapping bindings. The local label binding table and associated VERSION ID must be maintained across a LDP session restart. This version id can be managed either on a router basis or on a per session level and is left to the implementer.

The LMR similarly, will need to keep 1) bindings learned from an LMS in a remote binding table, and 2) the last VERSION ID "assign" value learned from the LMS. Both the remote binding table and the LMS version ID "assign" value needs to be maintained across a LDP session restart.

After session establishments, the LMS must wait for a VERSION ID "request" message from the LMR. The LMR sends the VERSION ID that it last processed from the LMS. If the LMR needs to purge its remote binding table, it can optionally send a VERSION ID=0 "request" to the LMS request that it re-send all its bindings. The LMR does not necessarily send a VERSION-ID "request" TLV immediately after a session is established. It sends it the request when it's ready to receive bindings.

After receiving a VERSION-ID "request" message from the LMR, a LMS should send a VERSION_ID "assign" message starting with version ID not greater than the VERSION-ID requested by the LMR. The LMS should then scan it's local label binding table and advertise/withdraw bindings with version id's starting with the version id it indicated in the VERSION-ID "assign" message. The LMS must also scan its local label binding table and mark all previously withdrawn bindings with VERSION-ID less than the "request" version ID as released.

If the LMS is not able to honor sending with the requested Version-ID, it should send a VERSION-ID "assign" message with VERSION-ID equal to zero to indicate that all previously advertised label bindings should be discarded and that the LMS will be re-advertising all bindings. The bindings to be removed needs to be marked stale and purged if not reclaimed after the GR recovery period.

Unlike the existing LDP Graceful restart behavior, after a session goes down and is re-established, label bindings previously advertised are not implied withdrawn, so a LMS MUST keep track of all its bindings and withdraw them. If bindings are deleted from the local label binding table while a "downed" neighbor is still in a LDP GR recovery period, that session must be flagged to indicate that if the LDP session re-establishes, then a VERSION ID "assign" message must

use a value=0, to force the LMR to purge all previously learned bindings.

5.3. ILA Version ID=0 Handling

If a LMS sets this version ID:

- 1) the LMR must purge all its previously received label bindings as the LMS will not be sending label withdraw's for previously advertised label mappings.

- 2) The LMS does not need to send label withdraws for previously advertised bindings, as all previously advertised bindings are implied withdrawn.

5.4. ILA Version ID Assignment

Version ID's value can be incremented. The grouping of the number of label mapping updates per Version ID is up to the implementer. It's suggested that the default value is 100 label updates messages per unique Version ID. Each subsequent assigned value MUST be greater than the previously assigned value, but need not be contiguous. For example, the implementer may choose to have a unit Version id per update, but choose to send a Version ID "assign" message per 100 update. In this case, the LSR may send "assign" a Version id of 1, 101, 201,... The spacing of the increment is left to the implementor.

5.5. ILA Version ID wraparound

The Version ID field is defined as a 64 bit value so wraparound is unlikely. This section defines how to handle this rare case. Since the each subsequent version ID needs to be assigned a value greater than the previously assigned value, when a wrap around occurs, the LMS must reset the LDP session, update the Version ID assigned to every binding and send a Version ID "assign" message after the session re-establishes. Since LDP graceful restart is supported by the sessions that employ ILA, resetting the session will not disrupt forwarding as the existing LDP GR mechanism should protect traffic.

6. Acknowledgements

The authors wish to thank Bin Mo and Eric Rosen for their comments.

7. IANA Considerations

This draft require any new allocations by IANA for the 1) LDP ILA capability TLV, 2) LDP ILA Version ID TLV, and 3) LDP ILA Version Message..

8. Security Considerations

This extension to LDP does not change the underlying security issues inherent in the existing [RFC3478] and [RFC5036]

9. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC3478] Leelanivas, M., Rekhter, Y., and R. Aggarwal, "Graceful Restart Mechanism for Label Distribution Protocol", RFC 3478, February 2003.
- [RFC5036] Andersson, L., Minei, I., and B. Thomas, "LDP Specification", RFC 5036, October 2007.
- [RFC5561] Thomas, B., Raza, K., Aggarwal, S., Aggarwal, R., and J. Le Roux, "LDP Capabilities", RFC 5561, July 2009.

Authors' Addresses

Alton Lo
Cisco Systems
170 W. Tasman Drive
San Jose, CA 95134
USA

Email: altonlo@cisco.com

Keyur Patel
Cisco Systems
170 W. Tasman Drive
San Jose, CA 95134
USA

Email: keyupate@cisco.com

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

Email: vlim@cisco.com

Internet Engineering Task Force
Internet-Draft
Intended status: Standards Track
Expires: January 9, 2012

L. Li
L. Huang
China Mobile
N. So
Verison Business
A. Kvalbein
Resiliens Communication AS
B. zhang
Telus Communications
July 8, 2011

MPLS Multiple Topology Applicability and Requirements
draft-li-mps-mt-applicability-requirement-02

Abstract

This document describes the applicability and requirements for Multiprotocol Label Switching Multiple Topology (MPLS-MT). The applicability and requirements are presented from different angles. They are expressed from a customer's point of view, a service provider's point of view and a vendor's point of view.

Status of this Memo

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

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

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

This Internet-Draft will expire on January 9, 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. Terminology	4
2. Introduction	4
3. Applicability	6
3.1. Simplified Data-plane	6
3.2. Automation of inter-layer interworking	7
3.3. Migration without service disruption	7
3.4. Protection using MT	8
3.5. Service Separation	8
3.6. Load Balancing	8
3.7. Inter-domain MPLS-TE and MPLS VPN	9
3.8. IPv6 deployment in IPv4 backbone	9
4. Service requirements	9
4.1. Availability	10
4.1.1. Physical Diversity and FRR	10
4.2. Stability	10
4.3. Traffic types	10
4.4. Data isolation	11
4.5. Security	11
4.5.1. User data security	11
4.5.2. Access control	11
4.5.3. MT router authentication and authorization	12
4.5.4. Inter domain security	12
4.6. Topology	12
4.7. Addressing	12
4.8. Quality of Service	13
4.9. Network Resource Partitioning and Sharing between MPLS-MTs (REWRITE with emphasis/focus on partition)	13
5. Provider requirements	13
5.1. Scalability	13
5.1.1. Service Provider Capacity Sizing Projections	14
5.1.2. MPLS-MT Scalability aspects	14
5.1.3. Number of MPLS-MTs in the network	14
5.1.4. Number of MPLS-MTs per customer	15
5.1.5. Number of addresses and address prefixes per MPLS-MT	15
5.1.6. Solution-Specific Metrics	15
5.2. Management	15
5.3. Customer Management of a MPLS-MT	16

- 6. Engineering requirements 16
 - 6.1. Forwarding plane requirements 16
 - 6.2. Control plane requirements 16
 - 6.3. Control Plane Containment 17
 - 6.4. Requirements for commonality of MPLS-MT mechanisms 17
 - 6.5. Interoperability 17
- 7. IANA Considerations 18
- 8. Acknowledgement 18
- 9. References 18
 - 9.1. Normative References 18
 - 9.2. Informative References 19
- Authors' Addresses 19

1. Terminology

Terminology used in this document

Non-MT: router Routers that do not have the MT capability.

MT router: Routers that have MT capability as described in this document.

MT-ID: Renamed TOS field in LSAs to represent Multi-Topology ID.

Default topology: Topology that is built using the TOS 0 metric (default metric).

MT topology: Topology that is built using the corresponding MT-ID metric.

MT: Shorthand notation for MPLS Virtual Topology.

MT#0 topology: Representation of TOS 0 metric in MT-ID format.

Non-MT-Area: An area that contains only non-MT routers.

MT-Area: An area that contains both non-MT routers and MT routers, or only MT routers.

2. Introduction

"Multi-Topology Routing in OSPF", RFC4915, describes a mechanism for Open Shortest Path First protocol to support Multi-Topologies (MTs) in IP network where the Type of Service (TOS) based metric fields are redefined and are used to advertise different topologies, each with a separate link metric. The classification of what type of traffic maps to which topology is not defined in RFC4915. The interface can be configured to belong to a set of topologies. Network topology changes will be advertised independently for each topology using a Multi-Topology Identifier (MT-ID), so that IP packets can be forwarded in the specific network topology independently.

"M-ISIS: Multi Topology (MT) Routing in Intermediate System to Intermediate Systems (IS-ISs)", RFC5120, describes a mechanism within Intermediate System to Intermediate Systems (IS-ISs) to run a set of independent IP topologies. The existing IS-IS protocol is extended so that the formation of adjacencies and advertising of prefixes and reachable intermediate system are performed independently within each topology.

There is a need to support Multiple-Topologies in MPLS network where a label switch path (LSP) is established within one topology or across multiple topologies and the traffic can be forwarded along the LSP within each network topology or across multiple topologies.

This document presents requirements for Multiprotocol Label Switching Multiple-Topology (MPLS-MT). It identifies requirements that may apply to one or more individual approaches that a Service Provider may use to provision LSPs in MPLS-MT. The specification of technical means to provide MPLS-MT services is outside the scope of this document. Other documents are intended to cover this aspect. This document is intended as a "checklist" of requirements, providing a consistent way to evaluate and document how well each approach satisfies specific requirements. The applicability statement documents for each approach should provide the results of this evaluation. This document is not intended to compare one approach to another. This document presents requirements from several points of view. It begins with some considerations from a point of view common to customers and service providers, continues with a customer perspective, and concludes with specific needs of a Service Provider (SP).

There are three different deployment scenarios MPLS-MT services are considered in this document:

1. Single-provider, single-AS: This is the least complex scenario, where the MPLS-MT service is offered across a single service provider network spanning a single Autonomous System.
2. Single-provider, multi-AS: In this scenario, a single provider may have multiple Autonomous Systems (e.g., a global Tier-1 ISP with different ASes depending on the global location, or an ISP that has been created by mergers and acquisitions of multiple networks). This scenario involves the constrained distribution of routing information across multiple Autonomous Systems.
3. Multi-provider: This scenario is the most complex, wherein trust negotiations need to be made across multiple service provider backbones in order to meet the security and service level agreements for the MPLS-MT customer. This scenario can be generalized to cover the Internet, which comprises of multiple service provider networks. It should be noted that customers can construct their own MPLS-MTs across multiple providers. However such MPLS-MTs are not considered here as they would not be "Providerprovisioned".

MPLS-MT is set of extensions to existing MPLS signaling protocols that makes MPLS signaling protocols aware of multi-topology. In the context of MPLS signaling the term "Multi-topology" is redefined to

be protocol independent unlike IGP-MT, which is scoped inside a single flavor IGP (ex. ISIS-MT or OSPF-MT). In other words, a MPLS Multiple-Topology can be mapped to a OSPFv2 based topology, OSPFv3 based topology or ISIS based IGP topology. Besides, a MPLS multi-topology can also be mapped to an instance of OSPF-MT or ISIS-MT. There are two major categories for MPLS-MT applications: a) MPLS RSVP-TE-MT applications, b) MPLS LDP-MT applications. The following sections of the draft describe application scenarios and MPLS-MT signaling in general. These application scenarios are useful for service providers who already have an MPLS network, or for service providers willing to migrate from IP to MPLS.

The following Sections describe applicability and generic MPLS-MT requirements.

3. Applicability

There are two main scenarios for how MPLS-MT can be used as a value-adding tool: 1) It can be exposed to and used by the customer to suit particular needs. For example, a customer might be given the option to select from a range of different topologies with different price and quality characteristics, and can select one (or more) that fulfills the given requirements. This could allow a service provider to better exploit network resources, by using pricing as an incentive. 2) It can be used as a management tool by the network operator to achieve certain goals such as resilience, traffic isolation and congestion avoidance, without exposing this to customers. Of course, one scenario does not exclude the other: an operator might want to offer MT routing to large customers, while also using it as a tool for "internal" purposes for its best effort services.

3.1. Simplified Data-plane

IGP-MT requires additional data-plane resources to maintain a separate forwarding table 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, the 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, and there is no need to have extra MT-FIBs. Forwarding IP packets over a particular MT requires either configuration or some external means at every node, to map an attribute of incoming IP packet header to IGP-MT, which is additional overhead for network management. With MPLS-MT, mapping is required only at the ingress-PE of an MPLS-MT LSP, because 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 a particular path (ERO) that can traverse different network layers 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 the default Layer-3 link-state topology as illustrated in the following diagram. With this mechanism a (G)MPLS-TE LSP can be offloaded to lower layers without service disruption and without complexity of configuration.

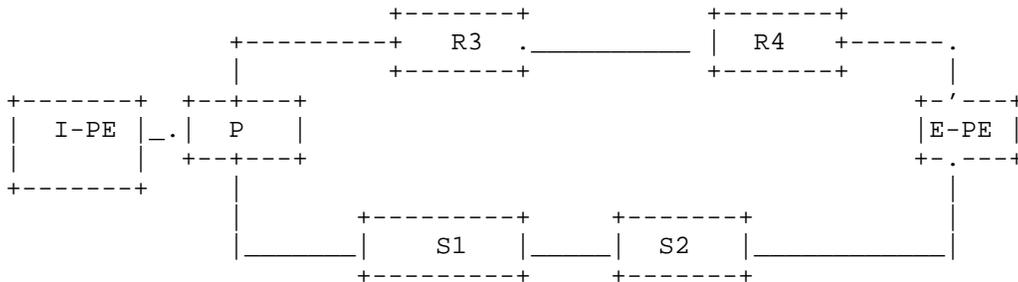


Figure 1: Layer-3 Link State Topology

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 stated above, MPLS-MT abstracts link state topology and identifies it by a unique MT-ID, which need not be the same as the IGP-MT ID. This characteristic is quite useful for service providers looking to migrate to a different flavor of IGP, e.g., OSPFv2 to ISIS6, OSPFv2 to OSPFv3. Service providers would like to incrementally upgrade their topologies, which requires an LSP to traverse multiple IGP

domains (OSPFv2 to OSPFv3) or (OSPF to ISIS). Migrating TE-LSPs to use a newly deployed link state topology requires a non-trivial effort. This migration may involve service disruption, especially when a path includes loose-hops in the ERO. For example: When an incoming PATH message requires an LSR to resolve loose-hop over a newly deployed IGP domain, which is not possible in the absence of MPLS-MT signaling. MPLS-MT allows an ingress-PE to specify Multiple-Topology to be used at every hop.

3.4. Protection using MT

We know that [IP-FRR-MT] can be used for configuring alternate paths via backup-mt, such that if the primary link fails, then a backup-MT can be used for forwarding. However, such techniques require special marking of IP packets that are forwarded using backup-MTs.

MPLS-LDP-MT procedures simplify the forwarding of the MPLS packets over backup-MTs, as the MPLS-LDP-MT procedure distributes separate labels for each MT. How backup paths are computed depends on the implementation, and the algorithm. MPLS-RSVP-MT in conjunction with IGP-MT could be used to separate the primary traffic and backup traffic. For example, service providers can create a backup MT that consists of links that are meant only for backup traffic. Service providers can then establish bypass LSPs, standby LSPs, using backup MT, thus keeping undeterministic backup traffic away from the primary traffic.

Another case is for mLDP. Since mLDP is based on IGP, in a same topology(or AS), when a backup mLDP is construct, it is diffult for the backup mLDP to find physical links different from those in the primary mLDP. With MPLS MT, the primary mLDP and backup mLDP can be construct in different topology with differnt links. So it is quite easy to set up primary and secondary mLDP with differnt links.

3.5. Service Separation

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

3.6. Load Balancing

MPLS-MT can be used to construct several alternative LSPs between PE routers. The LSPs in different topologies might follow partly overlapping routes through the network, or be completely disjoint. By smart assignment traffic to different MTs at the PE routers, it is

possible to offload traffic from heavily loaded links, and hence reduce the risk of congestion and improve resource utilization. This type of load balancing can be performed either in an offline way, where traffic is assigned to each MT according to a static split ratio, or in an online fashion, where the amount of traffic assigned to each MT according to a dynamic splitting function that depends on the current load situation.

3.7. Inter-domain MPLS-TE and MPLS VPN

Without MPLS MT, when a MPLS-TE tunnel is to be construct across multiple AS, PCE function has to be supported in each router in the MPLS-TE tunnel. With MPLS MT, routers in different ASs can be included into a new AS. In this way, MPLS-TE tunnels can be easily set up.

MPLS VPNs can be constructed across differnet ASs. In some case, when there are multiple ASs, it is quite difficult to manage the inter-as MPLS VPN. With MPLS MT, all the PEs can be configured in a single AS, so the MPLS VPN can be constructed and managed easily.

3.8. IPv6 deployment in IPv4 backbone

Without MPLS MT, the backbone is a sole topology to support IPv4 and IPv6 applications, IPv6 traffic is encoded into MPLS forwarding and mix with IPv4 traffic regardless using 6PE or 6VPN transition technology. As a result, service providers lose the visibility of traffic distribution breaking down to various protocols level which lead to impossible of accurately forecasting and planning of new IPv6 applications.

Besides, transition technology like 6PE or 6VPN request specific automatically generated IPv6 address to be used as next-hop and additional IPv6 routing and forwarding table to be maintained on gateway router, those tables coexist with IPv4 and consequentlly increase the amount of entire routes and bring challenge to enforce unified route and policy control. MPLS MT can help to de-couple the plane and simplify operation.

Obtaining a logical isolated end to end IPv6 plane via MPLS MT boosts the intergation of IPv6 exclusive application simulation and SLA measurement. It helps to capture the data on each hop and get a thorough view by hop by hop analyzing.

4. Service requirements

These are the requirements that a customer can observe or measure for

verifying whether the MPLS-MT service that the Service Provider (SP) provides is satisfactory. As mentioned before, each of these requirements apply equally across each of the three deployment scenarios unless stated otherwise.

4.1. Availability

MPLS-MT services **MUST** have high availability. LSPs that cross over several MTs require connectivity to be maintained even in the event of network failures.

This can be achieved via various redundancy techniques such as:

4.1.1. Physical Diversity and FRR

A single MT router may be connected to multiple MT routers. For a LSP, both local protections and global protections can be set up. Thus when a network failure happens, the traffic carried by the LSP can continue to flow across the MTs from the head end of the LSP to the tail ends of the LSP.

It should be noted that it is difficult to guarantee high availability when the MPLS-MT service is across multiple providers, unless there is a negotiation between the different service providers to maintain the service level agreement for the MPLS-MT customer.

4.2. Stability

In addition to availability, MPLS-MT services **MUST** also be stable. Stability is a function of several components such as MT routing and MPLS-MT signaling. For example, in the case of MT routing, route flapping or routing loops **MUST** be avoided in order to ensure stability. Stability of the MPLS-MT service is directly related to the stability of the mechanisms and protocols used to establish LSPs. It **SHOULD** also be possible to allow network upgrades and maintenance procedures without impacting the MPLS-MT service.

4.3. Traffic types

MPLS-MT services **MUST** support unicast (or point to point) traffic and **SHOULD** support multicast (or point-to-multipoint) traffic. For multicast traffic, the network delivers a stream to a set of destinations that have registered interest in the stream through a P2MP LSP. It is desirable to support multicast limited in scope to an intranet or extranet. The solution **SHOULD** be able to support a large number of such intranet or extranet specific multicast groups in a scalable manner. All MPLS-MT approaches **SHALL** support both IPv4 and IPv6 traffic.

4.4. Data isolation

The MPLS-MT MUST support forwarding plane isolation. The network MUST never deliver user data across MPLS-MT boundaries unless the two MPLS-MTs participate in an intranet or extranet.

Furthermore, if the provider network receives signaling or routing information from one MPLS-MT, it MUST NOT reveal that information to another MPLS-MT unless the two MPLS-MTs participate in an intranet or extranet. It should be noted that the disclosure of any signaling/routing information across an extranet MUST be filtered per the extranet agreement between the organizations participating in the extranet.

4.5. Security

A range of security features SHOULD be supported by the suite of MPLS-MT solutions in the form of securing customer flows, providing authentication services for temporary, remote or mobile users, and the need to protect service provider resources involved in supporting a MPLS-MT. Each MPLS-MT solution SHOULD state which security features it supports and how such features can be configured on a per customer basis. Protection against Denial of Service (DoS) attacks is a key component of security mechanisms.

Some security mechanisms may be equally useful regardless of the scope of the MPLS-MT. Other mechanisms may be more applicable in some scopes than in others. For example, in some cases of single-provider single-AS MPLS-MTs, the MPLS-MT service may be isolated from some forms of attack by isolating the infrastructure used for supporting MPLS-MTs from the infrastructure used for other services. However, the requirements for security are common regardless of the scope of the MPLS-MT service.

4.5.1. User data security

MPLS-MT solutions that support user data security SHOULD use standard methods to achieve confidentiality, integrity, authentication and replay attack prevention. Such security methods MUST be configurable between different end points. It is also desirable to configure security on a per-LSP basis. User data security using encryption is especially desirable in the multi-provider scenario.

4.5.2. Access control

A MPLS-MT solution may also have the ability to activate the appropriate filtering capabilities upon request of a customer. A filter provides a mechanism so that access control can be invoked at

the point(s) of communication between different organizations involved in an extranet. Access control can be implemented by a firewall, access control lists on routers, cryptographic mechanisms or similar mechanisms to apply policy-based access control. Such access control mechanisms are desirable in the multi-provider scenario.

4.5.3. MT router authentication and authorization

A MPLS-MT solution requires authentication and authorization of the following:

1. temporary and permanent access for users connecting to a MT router (authentication and authorization BY the MT router)
2. the MT router itself (authentication and authorization FOR the MT router)

4.5.4. Inter domain security

The MPLS-MT solution MUST have appropriate security mechanisms to prevent the different kinds of Distributed Denial of Service (DDoS) attacks, misconfiguration or unauthorized accesses in inter domain MPLS-MT connections. This is particularly important for multiservice provider deployment scenarios. However, this will also be important in single-provider multi-AS scenarios.

4.6. Topology

An MPLS-MT implementation SHOULD support arbitrary, customer -defined connectivity to the extent possible, for example, from partial mesh to full mesh topology. These can actually be different from the topology used by the service provider. The MPLS-MT services SHOULD be independent of MPLS-MT technology. To the extent possible, a MPLS-MT service SHOULD be independent of the geographic extent of the deployment. Multiple MPLS-MTs per customer SHOULD be supported without requiring additional hardware resources.

4.7. Addressing

Each customer resource MUST be identified by an address that is unique within its MPLS-MT. It need not be identified by a globally unique address. Support for IPv4 private addresses as described in [RFC1918] and unique local IPv6 addresses as described in [RFC 4193] , as well as overlapping customer addresses SHALL be supported. One or more MPLS-MTs for each customer can be built over the same infrastructure without requiring any of them to renumber. The solution MUST NOT use NAT on the customer traffic to achieve that

goal. Interconnection of two networks with overlapping IP addresses is outside the scope of this document.

4.8. Quality of Service

A technical approach for supporting MPLS-MTs SHALL be able to support QoS via IETF standardized mechanisms such as Diffserv. Support for best-effort traffic SHALL be mandatory for all MPLS-MT types. The extent to which any specific MPLS-MT service will support QoS is up to the service provider. In many cases single-provider single-AS MPLS-MTs will offer QoS guarantees. Support of QoS guarantees in the multiservice-provider case will require cooperation between the various service providers involved in offering the service.

4.9. Network Resource Partitioning and Sharing between MPLS-MTs (REWRITE with emphasis/focus on partition)

Network resources such as memory space, FIB table, bandwidth and CPU processing SHALL be shared between MPLS-MTs and, where applicable, with non-MPLS-MT Internet traffic. Mechanisms SHOULD be provided to prevent any specific MPLS-MT from taking up available network resources and causing others to fail. SLAs to this effect SHOULD be provided to the customer. Similarly, resources used for control plane mechanisms are also shared. When the service provider's control plane is used to distribute MPLS-MT specific information and provide other control mechanisms for MPLS-MTs, there SHALL be mechanisms to ensure that control plane performance is not degraded below acceptable limits when scaling the MPLS-MT service, or during network events such as failure, routing instabilities etc. Since a service provider's network would also be used to provide Internet service, in addition to MPLS-MTs, mechanisms to ensure the stable operation of Internet services and other MPLS-MTs SHALL be made in order to avoid adverse effects of resource hogging by large MPLS-MT customers.

5. Provider requirements

This section describes operational requirements for a cost-effective, profitable MPLS-MT service offering.

5.1. Scalability

The scalability for MPLS-MT solutions has many aspects. The list below is intended to comprise of the aspects that MPLS-MT solutions SHOULD address. Clearly these aspects in absolute figures are very different for different types of MPLS-MTs. It is also important to verify that MPLS-MT solutions not only scales on the high end, but

also on the low end - i.e., a MPLS-MT with three nodes and three users should be as viable as a MPLS-MT with hundreds of nodes and thousands of users.

5.1.1. Service Provider Capacity Sizing Projections

A MPLS-MT solution SHOULD be scalable to support a large number of MPLS-MTs per Service Provider network.

A MPLS-MT solution SHOULD be scalable to support of a large number of routes per MPLS-MT. The number of routes per MPLS-MT may range from just a few to ($O(10^5)$) exchanged between ISPs, with typical values being in the $O(10^3)$ range. The high end number is especially true considering the fact that many large ISPs may provide MPLS-MT services to smaller ISPs or large corporations.

A MPLS-MT solution SHOULD support high values of the frequency of configuration setup and change. Approaches SHOULD articulate scaling and performance limits for more complex deployment scenarios, such as single-provider multi-AS MPLS-MTs, multi-provider MPLS-MTs. Approaches SHOULD also describe other dimensions of interest, such as capacity requirements or limits, number of interworking instances supported as well as any scalability implications on management systems. A MPLS-MT solution SHOULD support a large number of customer interfaces on a single PE or CE with current Internet protocols.

5.1.2. MPLS-MT Scalability aspects

This section describes the metrics for scaling MPLS-MT solutions. These numbers are only representative and different service providers may have different requirements for scaling. Further discussion on service provider sizing projections is in Section 5.1.1. It should also be noted that the numbers given below would be different depending on whether the scope of the MPLS-MT is single-provider single-AS, single-provider multi-AS, or multiprovider. Clearly, the larger the scope, the larger the numbers that may need to be supported. However, this also means more management issues. The numbers below may be treated as representative of the single-provider case.

5.1.3. Number of MPLS-MTs in the network

The number of MPLS-MTs SHOULD scale linearly with the size of the access network and with the number of PEs. The number of MPLS-MTs in the network SHOULD be $O(10)$. This requirement also effectively places a requirement on the number of tunnels that SHOULD be supported in the network.

5.1.4. Number of MPLS-MTs per customer

In some cases a service provider may support multiple MPLS-MTs for the same customer of that service provider. For example, this may occur due to differences in services offered per MPLS-MT (e.g., different QoS, security levels, or reachability) as well as due to the presence of multiple workgroups per customer. It is possible that one customer will run up to $O(10)$ MPLS-MTs.

5.1.5. Number of addresses and address prefixes per MPLS-MT

Since any MPLS-MT solution SHALL support private customer addresses, the number of addresses and address prefixes are important in evaluating the scaling requirements. The number of address prefixes used in routing protocols and in forwarding tables specific to the MPLS-MT needs to scale from very few (for smaller customers) to very large numbers seen in typical Service Provider backbones. The high end is especially true considering that many Tier 1 SPs may provide MPLS-MT services to Tier 2 SPs or to large corporations. This number would be on the order of addresses supported in typical native backbones.

5.1.6. Solution-Specific Metrics

Each MPLS-MT solution SHALL document its scalability characteristics in quantitative terms. A MPLS-MT solution SHOULD quantify the amount of state that a PE and P device has to support. This SHOULD be stated in terms of the order of magnitude of the number of MPLS-MTs supported by the service provider.

5.2. Management

A service provider MUST have a means to view the topology, operational state, service order status, and other parameters associated with each customer's MPLS-MT. Furthermore, the service provider MUST have a means to view the underlying logical and physical topology, operational state, provisioning status, and other parameters associated with the equipment providing the MPLS -MT service(s) to its customers.

In the multi-provider scenario, it is unlikely that participating providers would provide each other a view to the network topology and other parameters mentioned above. However, each provider MUST ensure via management of their own networks that the overall MPLS -MT service offered to the customers are properly managed. In general the support of a single MPLS-MT spanning multiple service providers requires close cooperation between the service providers. One aspect of this cooperation involves agreement on what information about the

MPLS-MT will be visible across providers, and what network management protocols will be used between providers. MPLS-MT devices SHOULD provide standards-based management interfaces wherever feasible.

5.3. Customer Management of a MPLS-MT

A customer SHOULD have a means to view the topology, operational state, service order status, and other parameters associated with his or her MPLS-MT.

A customer SHOULD be able to make dynamic requests for changes to traffic parameters. A customer SHOULD be able to receive real-time response from the SP network in response to these requests. One example of such service is a "Dynamic Bandwidth management" capability, that enables real-time response to customer requests for changes of allocated bandwidth allocated to their MPLS-MT(s). A possible outcome of giving customers such capabilities is Denial of Service attacks on other MPLS-MT customers or Internet users. This possibility is documented in the Security Considerations section.

6. Engineering requirements

These requirements are driven by implementation characteristics that make service and provider requirements achievable.

6.1. Forwarding plane requirements

The SP is REQUIRED to provide per-MPLS-MT management, tunnel maintenance and other maintenance required in order to meet the SLA/SLS.

By definition, MPLS-MT traffic SHOULD be segregated from each other, and from non-MPLS-MT traffic in the network. After all, MPLS-MTs are a means of dividing a physical network into several logical or physical networks. MPLS-MT traffic separation SHOULD be done in a scalable fashion. However, safeguards SHOULD be made available against misbehaving MPLS-MTs to not affect the network and other MPLS-MTs.

A MPLS-MT solution SHOULD NOT impose any hard limit on the number of MPLS-MTs provided in the network.

6.2. Control plane requirements

The plug and play feature of a MPLS-MT solution with minimum configuration requirements is an important consideration. The MPLS-MT solutions SHOULD have mechanisms for protection against

customer interface and/or routing instabilities so that they do not impact other customers' services or impact general Internet traffic handling in any way.

A MPLS-MT SHOULD be provisioned with minimum number of steps. For this to be accomplished, an auto-configuration and an auto-discovery protocol, which SHOULD be as common as possible to all MPLS-MT solutions, SHOULD be defined. However, these mechanisms SHOULD NOT adversely affect the cost, scalability or stability of a service by being overly complex, or by increasing layers in the protocol stack.

Mechanisms to protect the SP network from effects of misconfiguration of MPLS-MTs SHOULD be provided. This is especially of importance in the multi-provider case, where misconfiguration could possibly impact more than one network.

6.3. Control Plane Containment

The MPLS-MT control plane MUST include a mechanism through which the service provider can filter MPLS-MT related control plane information as it passes between Autonomous Systems. For example, if a service provider supports a MPLS-MT offering, but the service provider's neighbors do not participate in that offering, the service provider SHOULD NOT leak MPLS-MT control information into neighboring networks. Neighboring networks MUST be equipped with mechanisms that filter this information should the service provider leak it. This is important in the case of multi-provider MPLS-MTs as well as singleprovider multi-AS MPLS-MTs.

6.4. Requirements for commonality of MPLS-MT mechanisms

The mechanisms used to establish a MPLS-MT service SHOULD re-use well-known IETF protocols as much as possible. It should, however, be noted that the use of Internet mechanisms for the establishment and running of an Internet-based MPLS-MT service, SHALL NOT affect the stability, robustness, and scalability of the Internet or Internet services. In other words, these mechanisms SHOULD NOT conflict with the architectural principles of the Internet, nor SHOULD it put at risk the existing Internet systems.

In addition to commonality with generic Internet mechanisms, infrastructure mechanisms used in different MPLS-MT solutions SHOULD be as common as possible.

6.5. Interoperability

Each technical solution is expected to be based on interoperable Internet standards.

Multi-vendor interoperability at network element, network and service levels among different implementations of the same technical solution SHOULD be ensured (that will likely rely on the completeness of the corresponding standard). This is a central requirement for SPs and customers.

The technical solution MUST be multi-vendor interoperable not only within the SP network infrastructure, but also with the customer's network equipment and services making usage of the MPLS-MT service.

Inter-domain interoperability - It SHOULD be possible to deploy a MPLS-MT solution across domains, Autonomous Systems, or the Internet.

7. IANA Considerations

TBD

8. Acknowledgement

Thanks for the contributions from Quintin Zhao, Ravi Tori, Huaimo Chen, Luyuang Fang, Chao Zhou.

9. References

9.1. Normative References

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

9.2. Informative References

Authors' Addresses

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

Ning So
Verison Business
2400 North Glenville Drive
Richardson, TX 78052
USA

Email: Ning.So@verizonbusiness.com

Amund Kvalbein
Resiliens Communication AS
Martin Linges v 17, Fornebu
Fornebu, Lysaker 1325
Norway

Email: Amundk@simula.com

Boris Zhang
Telus Communications
200 Consilium Pl Floor 15
Toronto, ON M1H 3J3
Canada

Email: Boris.Zhang@telus.com

Internet Engineering Task Force
Internet-Draft
Intended status: Informational
Expires: December 9, 2011

R. Martinotti
D. Caviglia
Ericsson
N. Sprecher
Nokia Siemens Networks
A. D'Alessandro
A. Capello
Telecom Italia
Y. Suemura
NEC Corporation of America
June 7, 2011

Interworking between MPLS-TP and IP/MPLS
draft-martinotti-mpls-tp-interworking-02

Abstract

Purpose of this ID is to illustrate interworking scenarios between network(s) supporting MPLS-TP and network(s) supporting IP/MPLS. Main interworking aspects, issues and open points are highlighted.

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 December 9, 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	3
1.1. Scope of this document	3
2. Conventions used in this document	3
3. Acronyms	3
4. Problem Statement	4
5. Terminology	4
6. Elements used in the figures	5
7. Interconnectivity Options	6
7.1. Network Layering model	8
7.1.1. OAM Implication of the Layering model	8
7.1.2. Layering model control plane consideration	8
7.2. Network Layering scenarios	8
7.2.1. Port based transparent transport of IP/MPLS	8
7.2.2. VLAN based transparent transport of IP/MPLS	12
7.2.3. Port based transport of IP/MPLS with Link Layer removal	12
7.2.4. IP/MPLS / MPLS-TP hybrid edge node	15
7.2.5. MPLS-TP carried over IP/MPLS	18
7.3. Network Partitioning Model	18
7.3.1. Connectivity constraints of the partitioning model	18
7.3.2. OAM Implications of the partitioning model	19
7.4. Network Partitioning scenarios	19
7.4.1. Border Node - Multisegment Pseudowire	20
7.4.2. Border Node - LSP stitching	22
7.4.3. Border Link - Multisegment Pseudowire	25
7.4.4. Border Link - LSP stitching	27
8. Acknowledgements	30
9. IANA Considerations	30
10. Contributing Authors	30
11. Security Considerations	32
12. References	32
12.1. Normative References	32
12.2. Informative References	32
Appendix A. Additional Stuff	32
Authors' Addresses	33

1. Introduction

1.1. Scope of this document

This document illustrates the most likely interworking scenarios between MPLS-TP and IP/MPLS. For each of the examined scenarios interworking aspects, limitations, issues and open points, with particular focus on OAM capabilities, are provided.

The main architectural construct considered in this document foresees PWE3 Protocol Stack Reference Model and MPLS Protocol Stack Reference Model. See [RFC 5921] for details.

2. Conventions used in this document

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

3. Acronyms

AC Attachment circuit
CE Customer Edge
CLI Client
CP Control Plane
DP Data Plane
ETH Ethernet MAC Layer
ETY Ethernet Physical Layer
IWF Interworking Function
LER Label Edge Router
LSP Label Switched Path
LSR Label Switch Router

MAC Media Access Control
MEP Maintenance Association End Point
MIP Maintenance Association Intermediate Point
MP Management Plane
MS-PW Multi Segment PW
NE Network Element
OAM Operations, Administration and Maintenance
PE Provider Edge
PHY Physical Layer
PSN Packet Switched Network
PW Pseudowire
SRV Server
SS-PW Single Segment PW
S-PE Switching Provider Edge
T-PE Terminating Provider Edge

4. Problem Statement

This document addresses interworking issues between MPLS-TP network and IP/MPLS network. The network decomposition can envisage network layering and/or network partitioning.

The presented scenarios are not intended to be comprehensive, for instance more complex scenarios can be created composing those described in this document.

5. Terminology

As far as this document is concerned, the following terminology is used:

- o IP/MPLS NE: a NE that supports IP/MPLS functions
- o IP/MPLS Network: a network in which IP/MPLS NEs are deployed
- o MPLS-TP NE: a NE that supports MPLS-TP functions
- o MPLS-TP Network: a network in which MPLS-TP NEs are deployed
- o Node: either MPLS-TP NE, IP/MPLS NE or CE
- o Ingress direction: from client to network
- o Egress direction: from network to client

For each of the scenarios described in this document, two paragraphs may appear, one related to possible issues already envisaged by the authors (Open Issues), the other related to aspects still left for further study and/or definition (Open Points).

This Section provides some terminology about network layering and partitioning. Primarily source of those definitions is [ITU-T G.805]. Readers already familiar with these concepts can skip this Section.

6. Elements used in the figures

A legenda of the symbols, which are most used in the following Sections, is provided, in order to facilitate comprehension of the scenarios.

```
Node:
----- Direct connection
- - - Virtual connection
..... one or more direct connections

Layers:
| Termination
+ Connection
<-> Stitching

OAM:
> or < MEP
O MIP
```

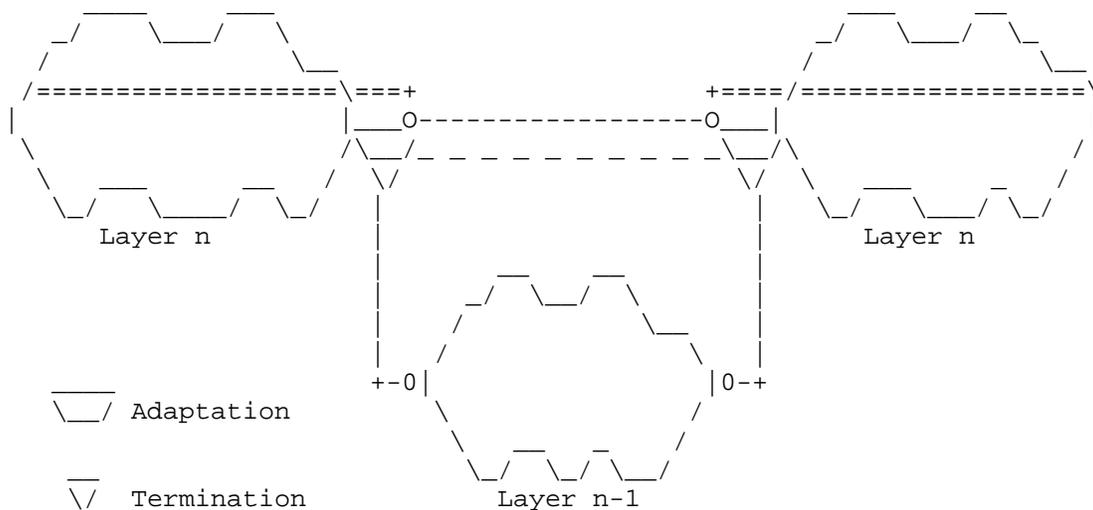
Figure 1

7. Interconnectivity Options

The MPLS-TP project adds dataplane OAM functionality to the MPLS tool set that permits executive action to be delegated to the dataplane. This provides the option of running MPLS without a control plane while still providing carrier grade resiliency options for connection oriented operation. Connection oriented operation alone does not offer the scalability to offer contemporary multipoint service solutions, but the combination of MPLS-TP connection oriented backhaul and IP/MPLS service capabilities permits the deployment of networks that scale significantly beyond the boundaries of current control plane scaling.

This section describes the methods in which IP/MPLS and MPLS-TP domains can interconnect. The network decomposition can envisage network layering and/or network partitioning. The presented scenarios are not intended to be comprehensive, for instance more complex scenarios can be created composing those described in this document. The various elements introduced in this section will be referred to in later sections.

The following figure illustrates the Network Layering concept, as it is described in Section 7.1:

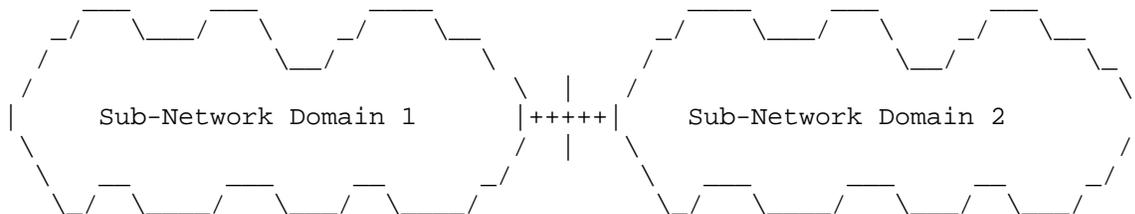


Network Layering

Figure 2

Layer n is carried over Layer n-1, via adaptation and termination functions. Some readers will also call this concept "Overlay model".

The following figure illustrates the Network Partitioning concept, as it is described in Section 7.3:



Network Partitioning

Figure 3

The boundary between the two subnetworks can be a link (as defined by [ITU-T G.805]), but also a Node, which in this case SHALL be able to

handle the technologies of both subnetworks.

The two subnetworks are at the same level. Some readers will also call this concept "Peer model".

7.1. Network Layering model

Two relationship are considered: the IP/MPLS network is carried over the MPLS-TP one, the MPLS-TP network is carried over the IP/MPLS one. This version of the draft focuses on the former relationship. In the MPLS-TP architecture, the pseudo wire is the primary unit of carriage of non-MPLS-TP payloads. This provides a clean demarcation between MPLS-TP operations and transported payloads.

7.1.1. OAM Implication of the Layering model

The overlay model has the virtue of uniform deployment of OAM capabilities and encapsulations at all MIPs and MEPs at a given layer in the label stack. The IP/MPLS architecture does include OAM transactions originated by MIPs so the layer interworking function for MPLS-TP servers is simplified.

7.1.2. Layering model control plane consideration

The interworking between an IP/MPLS domain and an MPLS-TP domain highly depends on the implemented model (i.e. layering or partitioning) and different scenarios can be implemented depending on a number of different aspects.

In the case of layering model, the first aspect consists on the provisioning of the LSP at the N-1 layer (MPLS-TP layer). Two possible scenarios are foreseen: pre-configuration of the MPLS-TP LSP or induced provisioning. The pre-configuration of the MPLS-TP LSP can be performed either manually via NMS or via the MPLS-TP control plane signaling and the MPLS-TP LSP can be exported to the IP/MPLS domain as a forwarding adjacency. On the other side the signaling messages at the IP/MPLS layer, upon reaching the border of the MPLS-TP domain, can induce the signaling of the MPLS-TP LSP via RSVP-TE. Other use cases depend on how the IP/MPLS is carried over the MPLS-TP domain and are analyzed scenario by scenario in the following sections.

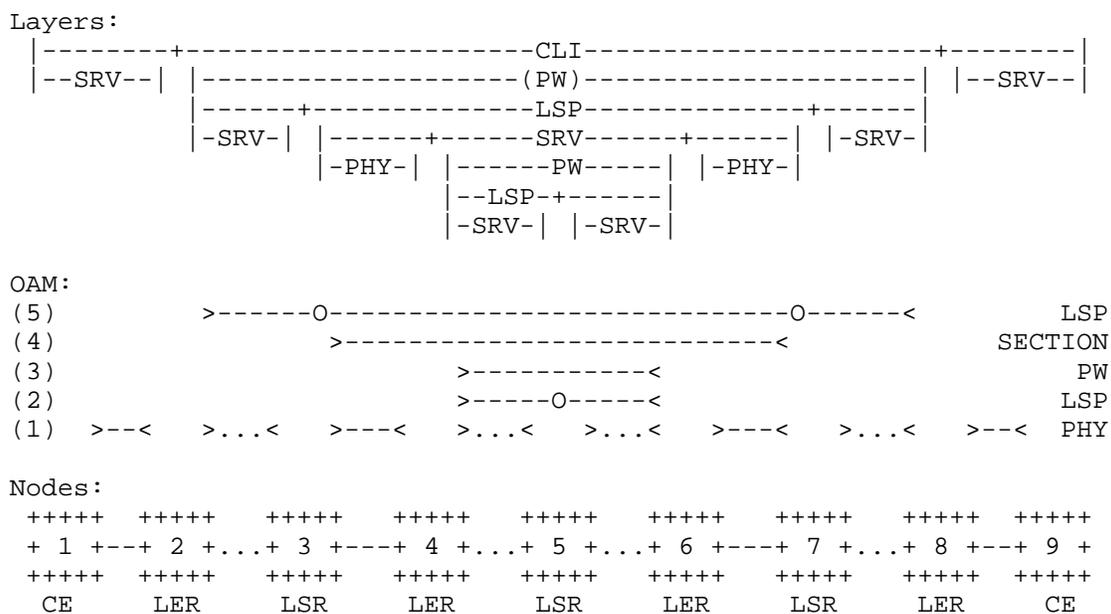
7.2. Network Layering scenarios

7.2.1. Port based transparent transport of IP/MPLS

This scenario foresees an IP/MPLS network carried over an MPLS-TP network. The selection of the route over the MPLS-TP network is done

7.2.1.1. OAM Considerations

The following figure illustrates the stacking relationship among the technology layers and OAM relationship among the networks:



Port based transparent transport - Layers and OAM view

Figure 5

Several levels of OAM are shown in the previous figure, these are not comprehensive and any subset of them MAY be configured in a network. A brief description of the different levels is provided:

- (5) Edge-to-Edge MPLS OAM on IP/MPLS network (at LSP level)
- (4) Section OAM on IP/MPLS network
- (3) Edge-to-Edge MPLS-TP OAM on MPLS-TP network (at PW level)
- (2) Edge-to-Edge MPLS-TP OAM on MPLS-TP network (at LSP level)

(1) Physical level OAM (MAY be of several kinds)

In case of fault detected at the MPLS-TP LSP (2) level, the corresponding server MEP asserts a signal fail condition and notifies that to the co-located MPLS-TP client/server adaptation function which then generates OAM packets with AIS information in the downstream direction to allow the suppression of secondary alarms at the MPLS-TP MEP in the client (sub-) layer, which in this example correspond to PW layer (3).

Note that the OAM layers not directly related to MPLS-TP network have been reported just for completeness of the scenario; however their behavior and interworking are out of scope of this document. For MPLS-TP Alarm reporting detailed description, please refer to [draft-ietf-mpls-tp-oam-framework].

7.2.1.2. Control Plane considerations

In this case the interconnection between the IP/MPLS domain and the MPLS-TP domain consists of a link. This does not allow a transparent transport of the IP control messages (e.g. LDP) over the MPLS-TP LSPs due to the fact that the egress node of the MPLS-TP domain is not able to route IP packets on its interfaces. The IP control messages need to be carried over an Ethernet frame over a PWE3 before being injected into the MPLS-TP LSP. In other words they are forwarded with two labels, the PWE3 one (S=0) and the LSP one (S=1). The IP control message, upon reaching the egress LER of the MPLS-TP domain, can be correctly forwarded to the ingress node of the IP/MPLS domain.

7.2.1.3. Services view

There are two service models supported by the overlay model when combined with Ethernet PWs. The first is simple p2p encapsulation and transport of all traffic presented to the MPLS-TP on a given interface. This is of limited utility due to the number of ports required to achieve the desired level of network interconnect across the MPLS-TP core.

The second is that the MPLS-TP LER maps VLANs to distinct PWs such that multiple IP/MPLS adjacencies can be supported over each interface between the IP/MPLS LSR and the MPLS-TP LER. This potentially can require a large number of IP/MPLS adjacencies overlaying the core.

In both cases the service can be unprotected or protected.

7.2.1.4. Resiliency considerations

In the scenario where the service is unprotected, resiliency is fully delegated to the IP/MPLS network, which will depend on a combination of routing convergence and/or FRR to maintain service. This will be at the expense of routing stability.

A protected service can offer significant improvements in routing stability with the exception that the link between the IP/MPLS LSRs and the MPLS-TP LERs and the MPLS-TP LERs themselves are single points of failure. There is an advantage in that the single points of failures are adjacent to the MPLS LSRs such that there is a high probability of such failures manifesting themselves immediately in the form of a physical layer loss-of-signal failure and thus accelerating recovery. Multiple failure scenarios may also result in the IP/MPLS overlay having to take action to recover connectivity but this would be gated by whatever OAM detection mechanisms were employed by the IP/MPLS layer as there is no equivalent of MPLS-TP LDI across the interconnect interface.

7.2.2. VLAN based transparent transport of IP/MPLS

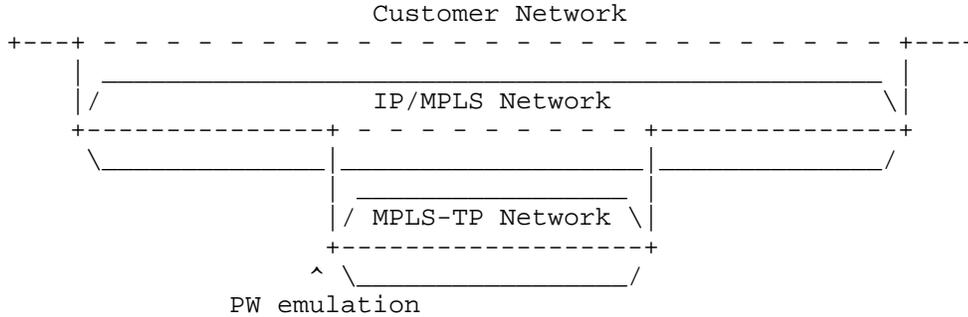
This scenario is analogous to the previous one. The interconnection between the IP/MPLS LSRs and the MPLS-TP PE is done via .1Q Tagged Ethernet, and VLANs are used to select the routes over the p2p Ethernet connectivity services over MPLS-TP (VPWS). The interworking is done via Ethernet encapsulation in PW over MPLS-TP (as per PWE3 Protocol Stack Reference Model). This VLAN based interconnection may be used in order to reduce the number of physical interfaces between the two networks. The same considerations of previous scenarios apply.

7.2.3. Port based transport of IP/MPLS with Link Layer removal

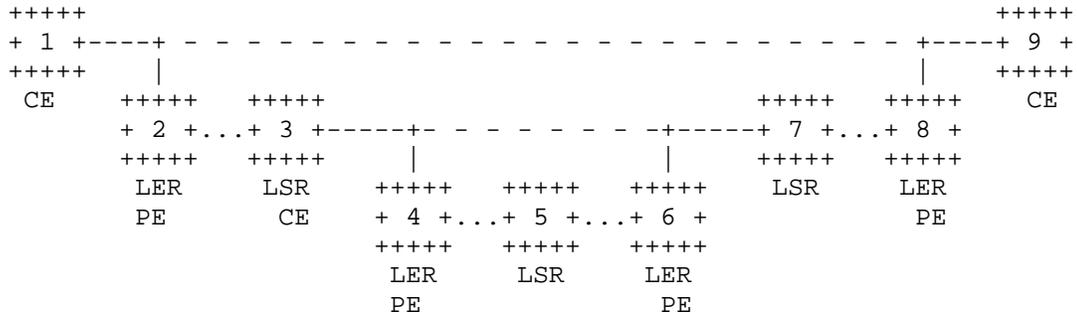
This scenario foresees an IP/MPLS network carried over an MPLS-TP network. The selection of the route over the MPLS-TP network is done on a per port basis. The physical interface between the IP/MPLS and the MPLS-TP network may be of different kind (e.g. Ethernet, POS); the interworking is done via Link Layer removal and client packet (MPLS and IP) encapsulation in PW over MPLS-TP (as per PWE3 Protocol Stack Reference Model). MPLS-TP LSPs are pre-configured with respect to IP/MPLS LSPs and are seen as routing adjacencies by the IP/MPLS network.

The following figure illustrates the functional interworking among the networks:

Networks:



Nodes:



Port based transport with Link Layer removal - Networks view

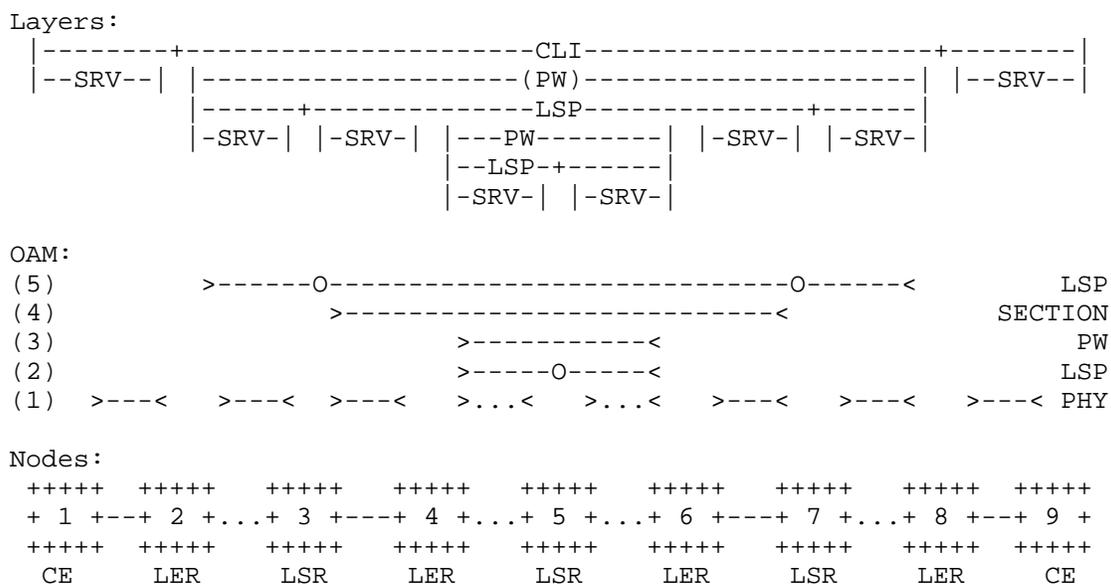
Figure 6

The LSR 3 and 7 are one hop away from the IP/MPLS layer point of view.

The service provided by the MPLS-TP network is p2p; client traffic is separated on a per port basis, so that (for example) all traffic coming from LSR 3 on the interface to LER 4 is transparently transported via LER 6 to LSR 7 and viceversa. The client traffic to be encapsulated is both MPLS packets (DP) and IP packets (DP, CP and MP). The encapsulation may be performed via PWs, that is, one PW is needed for MPLS and one for IP between any given port pair or directly using the LSP label stacking. The encapsulation via PW is required such that the IP/MPLS section preserves PHY like properties and to operationally isolate TP and IP/MPLS operation (e.g. reserved label handling link GAL and Router Alert).

7.2.3.1. OAM considerations

The following figure illustrates the stacking relationship among the technology layers and OAM relationship among the networks:



Port based transport with Link Layer removal - Layers and OAM view

Figure 7

Several levels of OAM are possible, a subset of them is shown in the previous figure, however these are not comprehensive, any subset of them MAY be configured in a network. A brief description of the levels is provided:

- (5) Edge-to-Edge MPLS OAM on IP/MPLS network (at LSP level)
- (4) Section MPLS OAM
- (3) Edge-to-Edge MPLS-TP OAM on MPLS-TP network (at PW level)
- (2) Edge-to-Edge MPLS-TP OAM on MPLS-TP network (at LSP level)

(1) Physical level OAM (MAY be of several kinds)

7.2.3.2. Control Plane considerations

In the case of transparent transport of the IP/MPLS over the MPLS-TP domain there are no differences, from a control plane point of view, with respect to the case of Ethernet encapsulation over MPLS-TP. Same considerations carried out in section 5.1.3.1.1.2 apply to this section.

7.2.3.3. Services view

The service model for the transparent transport mode is simple p2p encapsulation and transport of all traffic presented to the MPLS-TP on a given interface. This is of limited utility due to the number of ports required to achieve the desired level of network interconnect across the MPLS-TP core. It would potentially also require a correspondingly high number of IP/MPLS adjacencies to overlay the core.

The service can be unprotected or protected.

7.2.3.4. Resiliency considerations

The resiliency considerations are the same as for the overlay model.

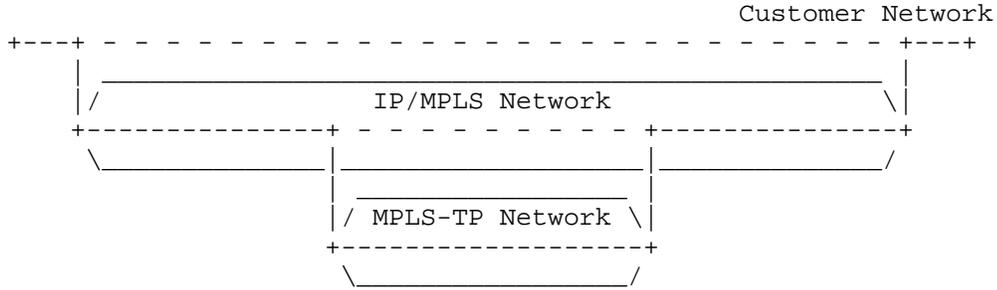
7.2.4. IP/MPLS / MPLS-TP hybrid edge node

In this scenario the physical interface between the IP/MPLS and the MPLS-TP network is generic and may be other than Ethernet (e.g. POS); the interworking is done via client LSP packet encapsulation as per MPLS labeled or IP traffic over MPLS-TP as per RFC 5921. MPLS-TP LSPs are pre-configured with respect to IP/MPLS LSPs and are seen as routing adjacencies between the hybrid edge nodes by the IP/MPLS network.

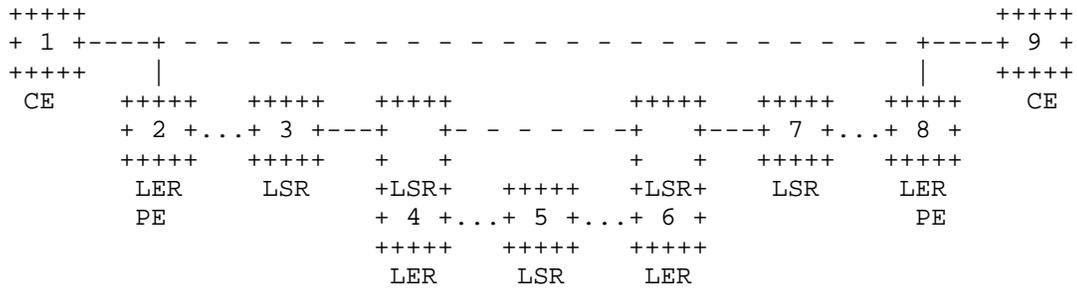
The service that is offered to the IP/MPLS network is that of a multi-point MPLS VPN.

The following figure illustrates the functional interworking among the networks.

Networks:



Nodes:



IP/MPLS encapsulation over MPLS-TP - Networks view

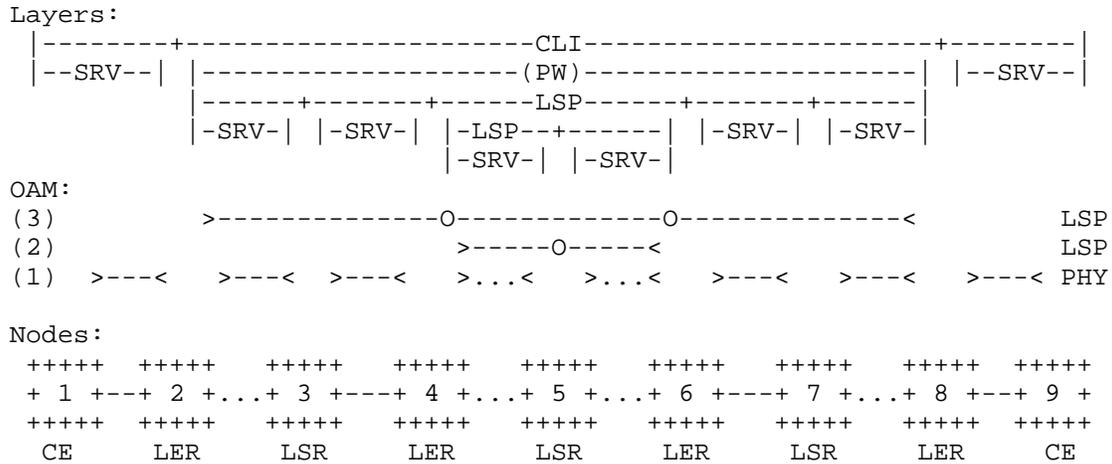
Figure 8

The Node 4 and 6 in the above figure act as dual function:

- o LSR of client IP/MPLS network
- o LER of server MPLS-TP subnetwork

7.2.4.1. OAM considerations

The following figure illustrates the stacking relationship among the technology layers and OAM relationship among the networks:



IP/MPLS encapsulation over MPLS-TP - Layers and OAM view

Figure 9

Several levels of OAM are possible, a subset of them is shown in the previous figure, however these are not comprehensive, any subset of them MAY be configured in a network. A brief description of the levels is provided:

- (3) Edge-to-Edge MPLS OAM on IP/MPLS network (at LSP level)
- (2) Edge-to-Edge MPLS-TP OAM on MPLS-TP network (at LSP level)
- (1) Physical level OAM (MAY be of several kinds)

7.2.4.2. Control Plane considerations

This case is different from the previous two because the interconnection between the IP/MPLS domain and the MPLS-TP domain consists of a node. This lead to the fact that IP control messages do not need to be carried over a PWE3 along the MPLS-TP domain but can be directly carried over an LSP. In other words they are forwarded with a single LSP label (S=1) and , upon reaching the hybrid node between the MPLS-TP domain and the next IP/MPLS domain , the signaling can be carried on.

7.2.4.3. Services view

The service model for the hybrid edge node model is that the MPLS-TP network appears to the IP/MPLS network as a complete IP/MPLS

subnetwork. This has the virtue of collapsing the number of IP/MPLS adjacencies required to overlay the core.

The service can be unprotected or protected. And the protection can be a combination of MPLS-TP resiliency and IP/MPLS recovery actions.

7.2.4.4. Resiliency considerations

The resiliency considerations are similar to that of the overlay model. However the extension of the control plane to the hybrid node means the lack of a dataplane LDI equivalent is mitigated, the IP/MPLS domain having been extended to reach the MPLS-TP OAM domain such that LDI indications from core failures can interwork directly the the control plane and accelerate recovery actions.

7.2.5. MPLS-TP carried over IP/MPLS

TODO

7.3. Network Partitioning Model

In the rest of this Section the following assumptions apply:

- o Customer network is carried partly over IP/MPLS subnetwork (e.g. via PW encapsulation) and partly over MPLS-TP subnetwork.
- o An example of server layer of MPLS is Ethernet.

For the purposes of this Section, MPLS-TP subnetwork is deployed between a CE and an IP/MPLS subnetwork. Other kinds of deployment are possible (not shown in this document), for instance:

- o More than two subnetworks are deployed between the CEs
- o MPLS-TP can be deployed between two subnetworks

7.3.1. Connectivity constraints of the partitioning model

The partitioning model is constrained to interconnecting LSPs or PWs with common behavioral characteristics. As MPLS-TP is constrained to connection oriented behavior the portion of the LSP that transits an IP/MPLS subnetwork will need to be effectively constrained to the same profile, that is connection oriented, and no PHP or merging. No ECMP or transit of LAG cannot be guaranteed which means OAM fate sharing may not exist in IP/MPLS subnetworks and the end-to-end OAM may only serve to coordinate dataplane resiliency actions between MEPs with respect to faults in the MPLS-TP subnetworks.

7.3.2. OAM Implications of the partitioning model

The partitioning model requires the concatenations of path segments that do not necessarily have common OAM components and have a number of possible implementations. At the simplest level configuration of common OAM capabilities and encapsulation between the MEPS in the MEG is required. The set that is common to the MEPS in the MEG may not necessarily be supported by the MIPs, and knowledge of MIP capability will not figure into MEP negotiation, so the MEPS may select a common mode that is not common with that supported by the MIPs.

The primary consequence being that MPLS-TP MIP originated transactions, or messages targeted to MIPs using MPLS-TP encapsulations will not be guaranteed to provide a uniform quality of information as not all MIPs will support MPLS-TP OAM extensions, and as noted will not participate in MEP-MEP configuration or negotiation.

This means that GAL encapsulated OAM may only serve to coordinate dataplane resiliency actions between MEPS with respect to faults in the MPLS-TP subnetworks and faults in the IP/MPLS subnetwork are recovered by IP/MPLS mechanisms (e.g. FRR). Edge to edge monitoring of MPLS/MPLS-TP networks may be implemented using an edge to edge LSP OAM/PW OAM, in order not to need a gateway/translation function on the border node between the two domains.

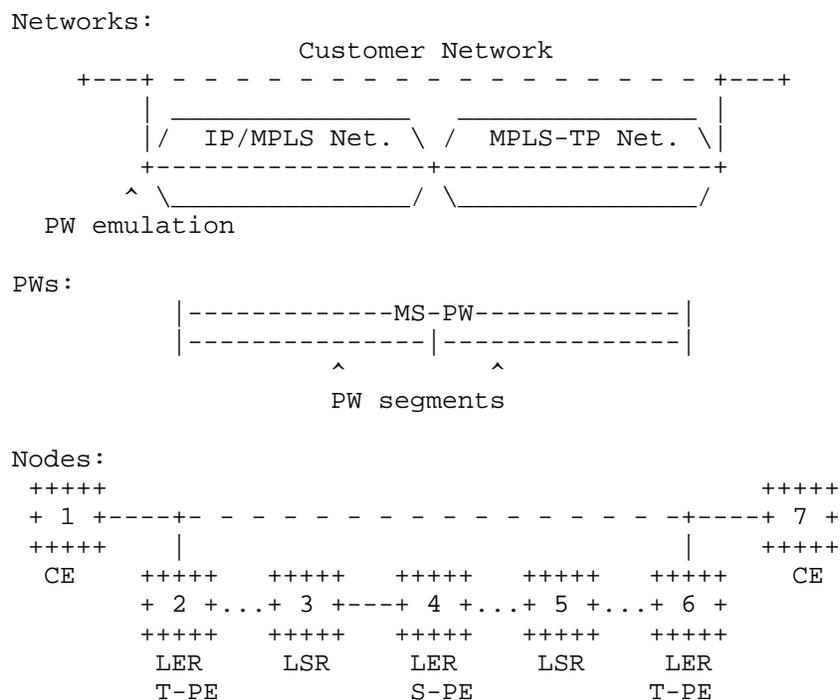
7.4. Network Partitioning scenarios

The main features to be taken into account in deploying a partitioned network are the following:

- o Border Node or Border Link
- o MultiSegment Pseudowire or LSP Stitching
- o Network Interworking
- o End-to-End OAM support
- o Interaction between DP of IP/MPLS and DP of MPLS-TP
- o Interaction between CP of IP/MPLS and MP of MPLS-TP
- o Interaction between CP of IP/MPLS and CP of MPLS-TP
- o Interaction between MP of IP/MPLS and MP of MPLS-TP

7.4.1. Border Node - Multisegment Pseudowire

The following figure illustrates the functional interworking among the networks:

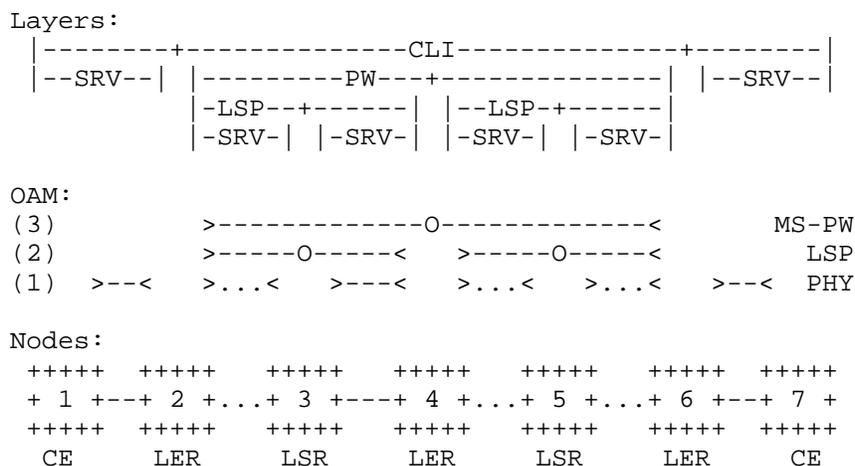


Border Node - Multisegment Pseudowire - Networks and PWs view

Figure 10

7.4.1.1. OAM considerations

The following figure illustrates the stacking relationship among the technology layers and OAM relationship among the networks:



Border Node - Multisegment Pseudowire - Layers and OAM view

Figure 11

Several levels of OAM are possible, a subset of them is shown in the previous figure, however these aren't comprehensive, any subset of them MAY be configured in a network. A brief description of the different levels is provided:

- (3) Edge-to-Edge MPLS/MPLS-TP OAM on whole network (at PW level)
- (2) Edge-to-Edge MPLS OAM and Edge-to-Edge MPLS-TP OAM on each network partition respectively (at LSP level)
- (1) Physical level OAM (MAY be of several kind)

Open Points:

- o Interworking between LSP OAM (2) and MS-PW OAM (3) is still to be cleared/defined
- o Edge-to-Edge MS-PW OAM (3) must be configured on different subnetworks

7.4.1.2. Control Plane considerations

TODO

7.4.1.3. Services view

The generalized service model for all partitioning models is a p2p connection for the PW client.

7.4.1.4. Resiliency considerations

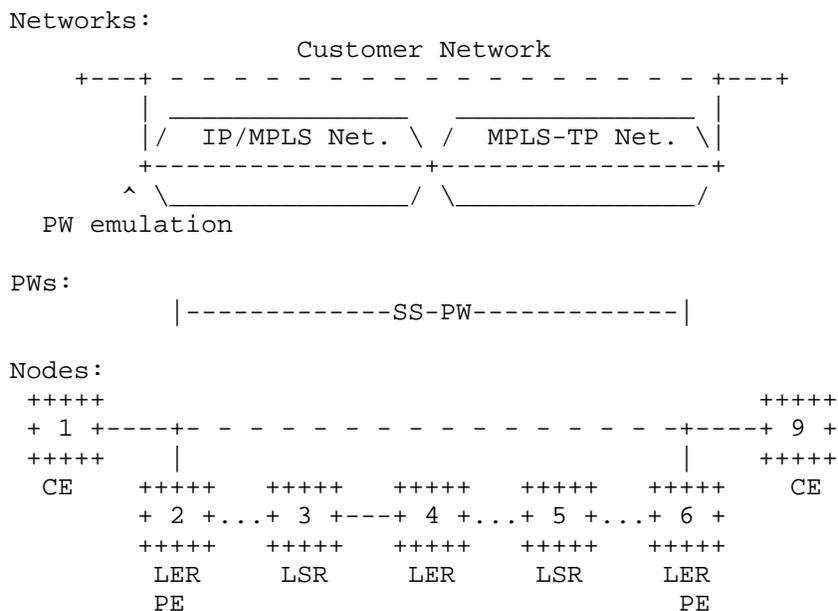
The PW can be configured to be protected or unprotected at the PW layer. If it is unprotected it is dependent on the underlying domains (MPLS-TP or IP/MPLS) resiliency mechanisms to offer subnetwork protection, but the S-PE is a single point of failure. A protected PW can be set up such that the working and protection PWs traverse physically diverse S-PEs.

Implementing E2E protection at the PW layer requires CC flows on the PW which for large numbers of PWs may have scaling implications.

When the PW is protected, the border node as an MS-PW stitching point permits the interworking of MPLS-TP fault indications with the PW signaling in the IP/MPLS domain such that fast E2E protection switching can be coordinated without requiring fast CC/CV OAM flows in the PW layer.

7.4.2. Border Node - LSP stitching

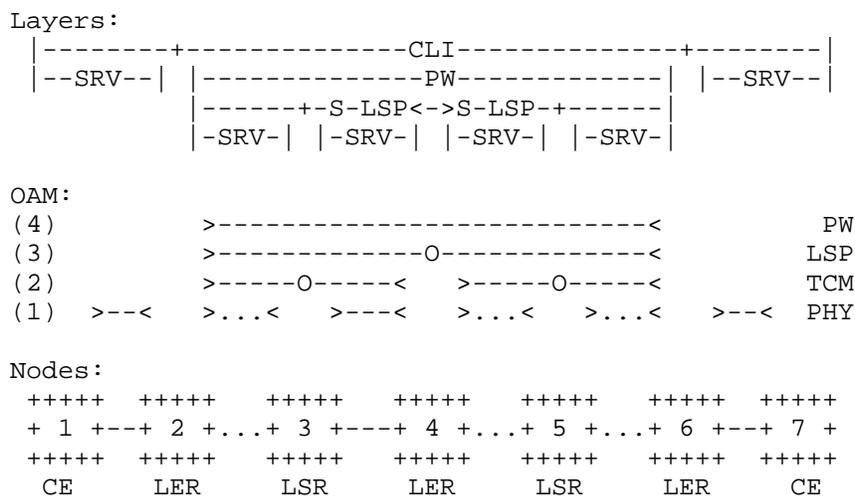
The following figure illustrates the functional interworking among the networks:



Border Node - LSP stitching - Networks and PWs view

Figure 12

The following figure illustrates the stacking relationship among the technology layers and OAM relationship among the networks:



Border Node - LSP stitching - Layers and OAM view

Figure 13

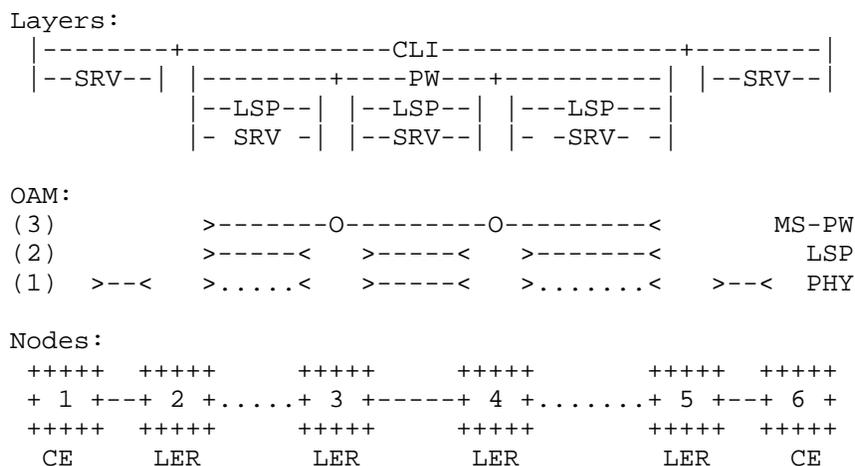
Note: in this case a SS-PW extends over the subnetworks as the stitched LSP does. TCM can be used to monitor the LSP segments.

Several levels of OAM are possible, a subset of them is shown in the previous figure, however these are not comprehensive, any subset of them MAY be configured in a network. A brief description of the different levels is provided:

- (4) Edge-to-Edge MPLS/MPLS-TP OAM on whole network (at PW level)
- (3) Edge-to-Edge MPLS/MPLS-TP OAM on whole network (at LSP level)
- (2) Edge-to-Edge MPLS OAM and Edge-toEdge MPLS-TP OAM on each network partition respectively (at TCM level)
- (1) Physical level OAM (MAY be of several kind)

Open Points:

- o Edge-to-Edge LSP OAM (3) must be configured on different subnetworks
- o Edge-to-Edge PW OAM (4) must be configured on different subnetworks



Border Link - Multisegment Pseudowire - Layers and OAM view

Figure 15

Several levels of OAM are possible, a subset of them is shown in the previous figure, however these are not comprehensive, any subset of them MAY be configured in a network. A brief description of the different levels is provided:

- (3) Edge-to-Edge MPLS/MPLS-TP OAM on whole network (at PW level)
- (2) Edge-to-Edge MPLS OAM, Border MPLS OAM and Edge-toEdge MPLS-TP OAM on each network partition respectively (at LSP level)
- (1) Physical level OAM (MAY be of several kinds)

Open Points:

- o Interworking between LSP OAM (2) and MS-PW OAM (3) is still to be cleared/defined
- o LSP between Node 3 and 4 could be avoided, however in this case PW over Ethernet should be specified.
- o Edge-to-Edge MS-PW OAM (3) must be configured on different subnetworks

7.4.3.2. Control Plane considerations

TODO

7.4.3.3. Services view

The generalized service model for all partitioning models is a p2p connection for the PW client.

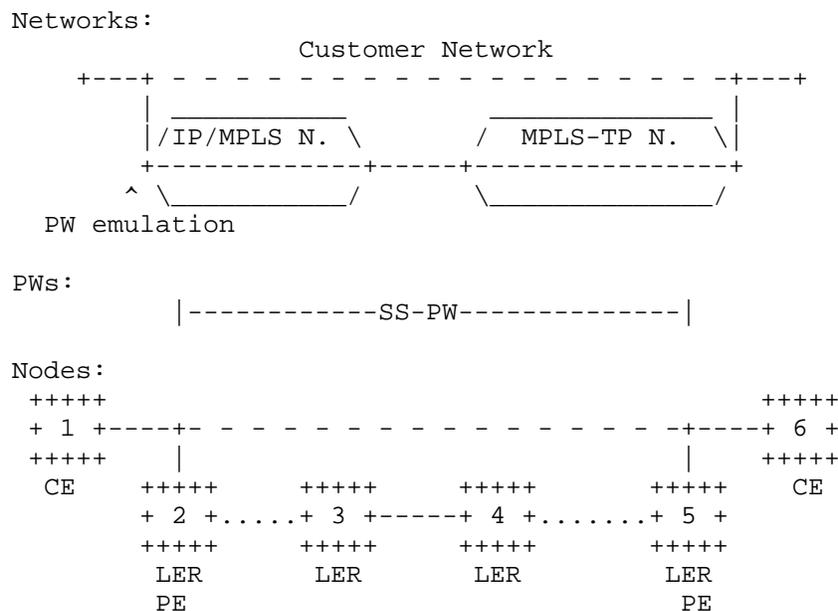
7.4.3.4. Resiliency considerations

The PW can be configured to be protected or unprotected at the PW layer. If it is unprotected it is dependent on the underlying domains (MPLS-TP or IP/MPLS) resiliency mechanisms to offer subnetwork protection, but the border S-PEs and border link are all single points of failure. A protected PW can be set up such that the working and protection PWs traverse physically diverse border links.

Implementing E2E protection at the PW layer requires CC flows on the PW which for largenumbers of PWs may have scaling implications.

7.4.4. Border Link - LSP stitching

The following figure illustrates the functional interworking among the networks:

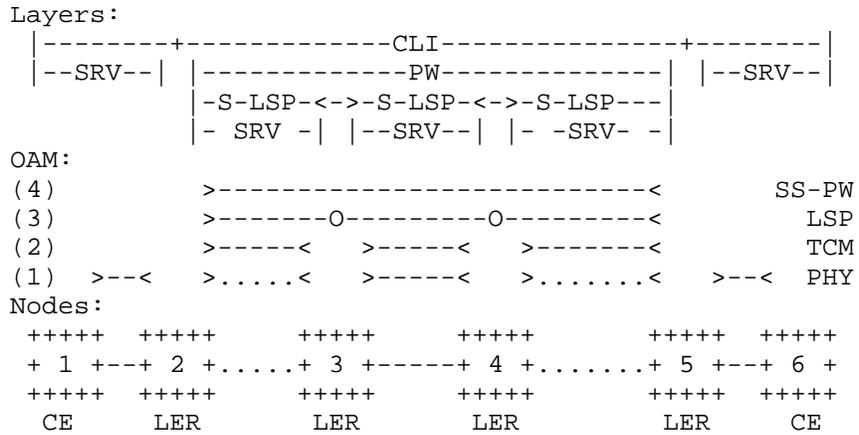


Border Link - LSP stitching - Networks view

Figure 16

7.4.4.1. OAM considerations

The following figure illustrates the stacking relationship among the technology layers and OAM relationship among the networks:



Border Link - LSP stitching - Layers and OAM view

Figure 17

Note: in this case a SS-PW extends over the subnetworks as the stitched LSP does. TCM can be used to monitor the LSP segments.

Several levels of OAM are possible, a subset of them is shown in the previous figure, however these aren't comprehensive, any subset of them MAY be configured in a network. A brief description of the different levels is provided:

- (4) Edge-to-Edge MPLS/MPLS-TP OAM on whole network (at PW level)
- (3) Edge-to-Edge MPLS/MPLS-TP OAM on whole network (at LSP level)
- (2) Edge-to-Edge MPLS OAM, Border MPLS OAM and Edge-to-Edge MPLS-TP OAM on each network partition respectively (at TCM level)
- (1) Physical level OAM (MAY be of several kinds)

7.4.4.2. Control Plane considerations

TODO

7.4.4.3. Services view

The generalized service model for all partitioning models is a p2p connection for the PW client.

7.4.4.4. Resiliency considerations

The LSP can be configured to be protected end to end, have subnetwork protection or be unprotected at the LSP layer. In the subnetwork protection scenario the border S-PEs and the borderlink are all single points of failure.

When GAL/GACH encapsulated OAM is deployed at (a minimum) of the LSP MEPs, it is possible to envision interworking of the MPLS-TP LSP and LSPs in the IP/MPLS domain set up with RSVP-TE and/or with LDP. In the latter case the MPLS-TP LSP maps to a FEC rather than a specific LSP but the MPLS_TP LSP would need to appear as a FEC in LDP with associated scaling impacts.

Open Points:

- o Edge-to-Edge LSP OAM (3) must be configured on different subnetworks
- o Edge-to-Edge PW OAM (4) must be configured on different subnetworks
- o Interworking between TCM OAM (2) and LSP OAM (3) is still to be cleared/defined
- o Interaction between IP/MPLS and MPLS-TP CPs is still to be cleared/defined

8. Acknowledgements

The authors gratefully acknowledge the input of Attila Takacs.

9. IANA Considerations

This memo includes no request to IANA.

10. Contributing Authors

David Allan
Ericsson
Holger Way
San Jose
U.S.

Email: david.i.allan@ericsson.com

Elisa Bellagamba
Ericsson
Torshamnsgatan 48
Stockholm 164 80
Sweden

Email: elisa.bellagamba@ericsson.com

Daniele Ceccarelli
Ericsson
Via A. Negrone 1/A
Genova - Sestri Ponente 16153
Italy

Email: daniele.ceccarelli@ericsson.com

David Saccon
Ericsson
Holger Way
San Jose
U.S.

Email: david.sacson@ericsson.com

John Volkering
Ericsson

Email: john.volkering@ericsson.com

11. Security Considerations

This document does not introduce any additional security aspects beyond those applicable to PWE3 and MPLS.

12. References

12.1. Normative References

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

12.2. Informative References

- [RFC3031] Rosen, E., Viswanathan, A., and R. Callon, "Multiprotocol Label Switching Architecture", RFC 3031, January 2001.
- [RFC5654] Niven-Jenkins, B., Brungard, D., Betts, M., Sprecher, N., and S. Ueno, "Requirements of an MPLS Transport Profile", RFC 5654, September 2009.
- [RFC5921] Bocci, M., Bryant, S., Frost, D., Levrau, L., and L. Berger, "A Framework for MPLS in Transport Networks", RFC 5921, July 2010.
- [RFC5860] Vigoureux, M., Ward, D., and M. Betts, "Requirements for Operations, Administration, and Maintenance (OAM) in MPLS Transport Networks", RFC 5860, May 2010.
- [RFC3209] Awduche, D., Berger, L., Gan, D., Li, T., Srinivasan, V., and G. Swallow, "RSVP-TE: Extensions to RSVP for LSP Tunnels", RFC 3209, December 2001.
- [RFC3473] Berger, L., "Generalized Multi-Protocol Label Switching (GMPLS) Signaling Resource ReserVation Protocol-Traffic Engineering (RSVP-TE) Extensions", RFC 3473, January 2003.
- [ITU-T G.805] "Generic functional architecture of transport networks", ID ITU-T G.805, March 2000.

Appendix A. Additional Stuff

This becomes an Appendix.

Authors' Addresses

Riccardo Martinotti
Ericsson
Via A. Negrone 1/A
Genova - Sestri Ponente 16153
Italy

Email: riccardo.martinotti@ericsson.com

Diego Caviglia
Ericsson
Via A. Negrone 1/A
Genova - Sestri Ponente 16153
Italy

Email: diego.caviglia@ericsson.com

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

Email: nurit.sprecher@nsn.com

Alessandro D'Alessandro
Telecom Italia
Via Reiss Romoli, 274
Torino 10148
Italy

Email: alessandro.dalessandro@telecomitalia.it

Alessandro Capello
Telecom Italia
Via Reiss Romoli, 274
Torino 10148
Italy

Email: alessandro.capello@telecomitalia.it

Yoshihiko Suemura
NEC Corporation of America
14040 Park Center Road
Herndon, VA 20171
USA

Email: Yoshihiko.Suemura@necam.com

MPLS Working Group
Internet Draft
Intended Status: Proposed Standard
Expires: November 2, 2011

Maria Napierala
AT&T

Eric C. Rosen
IJsbrands Wijnands
Cisco Systems, Inc.

May 2, 2011

Using LDP Multipoint Extensions on Targeted LDP Sessions

draft-napierala-mpls-targeted-mldp-01.txt

Abstract

Label Distribution Protocol (LDP) can be used to set up Point-to-Multipoint (P2MP) and Multipoint-to-Multipoint (MP2MP) Label Switched Paths. The existing specification for this functionality assumes that a pair of LDP neighbors are directly connected. However, the LDP base specification allows for the case where a pair of LDP neighbors are not directly connected; the LDP session between such a pair of neighbors is known as a "Targeted LDP" session. This document specifies the use of the LDP P2MP/MP2MP extensions over a Targeted LDP session.

Status of this Memo

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

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

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

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

Table of Contents

1	Introduction	3
1.1	Specification of requirements	3
1.2	Targeted mLDP	3
1.3	Targeted mLDP and the Upstream LSR	3
1.3.1	Selecting the Upstream LSR	3
1.3.2	Sending data from U to D	4
1.4	Applicability of Targeted mLDP	5
1.5	LDP Capabilities	5
2	Targeted mLDP with Unicast Replication	5
3	Targeted mLDP with Multicast Tunneling	6
4	IANA Considerations	8
5	Security Considerations	8
6	Acknowledgments	8
7	Authors' Addresses	8
8	Normative References	9

1. Introduction

1.1. 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].

1.2. Targeted mLDP

The Label Distribution Protocol (LDP) extensions for setting up Point-to-MultiPoint (P2MP) Label Switched Paths (LSPs) and Multipoint-to-Multipoint (MP2MP) LSPs are specified in [mLDP]. This set of extensions is generally known as "Multipoint LDP" (mLDP).

A pair of Label Switched Routers (LSRs) that are the endpoints of an LDP session are considered to be "LDP neighbors". When a pair of LDP neighbors are "directly connected" (e.g., they are connected by a layer 2 medium, or are otherwise considered to be neighbors by the a network's interior routing protocol), the LDP session is said to be a "directly connected" LDP session. When the pair of LDP neighbors are not directly connected, the session between them is said to be a "Targeted" LDP session.

The base specification for mLDP does not explicitly cover the case where the LDP multipoint extensions are used over a targeted LDP session. This document provides that specification.

We will use the term "Multipoint" to mean "either P2MP or MP2MP".

1.3. Targeted mLDP and the Upstream LSR

1.3.1. Selecting the Upstream LSR

In mLDP, a multipoint LSP (MP-LSP) has a unique identifier that is an ordered pair of the form <root, opaque value>. The first element of the ordered pair is the IP address of the MP-LSPs "root node". The second element of the ordered pair is an identifier that is unique in the context of the root node.

If LSR D is setting up the MP-LSP <R, X>, D must determine the "upstream LSR" for <R, X>. In [mLDP], the upstream LSR for <R, X>, U, is defined to be the "next hop" on D's path to R, and "next hop" is tacitly assumed to mean "IGP next hop". It is thus assumed that there is a direct LDP session between D and U. In this specification, we extend the notion of "upstream LSR" to cover the

following cases:

- U is the "BGP next hop" on D's path to R, where U and D are not IGP neighbors, and where there is a Targeted LDP session between U and D. In this case, we allow D to select U as the "upstream LSR" for <R,X>.
- If the "next hop interface" on D's path to R is an RSVP-TE P2P tunnel whose remote endpoint is U, and if there is known to be an RSVP-TE P2P tunnel from U to D, and if there is a Targeted LDP session between U and D, then we allow D to select U as the "upstream LSR" for <R,X>. This is useful when D and U are part of a network area that is fully meshed via RSVP-TE P2P tunnels.

The particular method used to select an "upstream LSR" is determined by the SP. Other methods than the ones above MAY be used.

1.3.2. Sending data from U to D

By using Targeted mLDP, we can construct an MP-LSP <R,X> containing an LSR U, where U has one or more downstream LSR neighbors (D1, ..., Dn) to which it is not directly connected. In order for a data packet to travel along this MP-LSP, U must have some way of transmitting the packet to D1, ..., Dn. We will cover two methods of transmission:

- Unicast Replication.

In this method, U creates n copies of the packet, and unicasts each copy to exactly one of D1, ..., Dn.

- Multicast tunneling.

In this method, U becomes the root node of a multicast tunnel, with D1, ..., Dn as leaf nodes. When a packet traveling along the MP-LSP <R,X> arrives at U, U transmits it through the multicast tunnel, and as a result it arrives at D1, ..., Dn.

When this method is used, it may be desirable to carry traffic of multiple MP-LSPs through a single multicast tunnel. We specify procedures that allow for the proper demultiplexing of the MP-LSPs at the leaf nodes of the multicast tunnel. We do not assume that all the leaf nodes of the tunnel are on all the MP-LSPs traveling through the tunnel; thus some of the tunnel leaf nodes may need to discard some of the packets received through the tunnel. For example, suppose MP-LSP <R1,X1> contains node U with downstream LSRs D1 and D2, while MP-LSP <R2,X2> contains node U

with downstream LSRs D2 and D3. Suppose also that there is a multicast tunnel with U as root and with D1, D2, and D3 as leaf nodes. U can aggregate both MP-LSPs in this one tunnel. However, D1 will have to discard packets that are traveling on <R2,X1>, while D3 will have to discard packets that are traveling on <R1,X2>.

1.4. Applicability of Targeted mLDP

When LSR D is setting up MP-LSP <R,X>, it MUST NOT use targeted mLDP unless D can select the "upstream LSR" for <R,X> using one of the procedures discussed in section 1.3.1.

Whether D uses Targeted mLDP when this condition holds is determined by provisioning, or by other methods that are outside the scope of this specification.

When Targeted mLDP is used, the choice between unicast replication and multicast tunneling is determined by provisioning, or by other methods that are outside the scope of this specification.

1.5. LDP Capabilities

Per [mLDP], any LSR that needs to set up an MP-LSP must support the procedures of [LDP-CAP], and in particular must send and receive the P2MP Capability and/or the MP2MP Capability. This specification does not define any new capabilities; the advertisement of the P2MP and/or MP2MP Capabilities on a Targeted LDP session means that the advertising LSR is capable of following the procedures of this document.

Some of the procedures of this document require the use of upstream-assigned labels [LDP-UP]. In order to use upstream-assigned labels as part of Targeted mLDP, an LSR must advertise the LDP Upstream-Assigned Label Capability [LDP-UP] on the Targeted LDP session.

2. Targeted mLDP with Unicast Replication

When unicast replication is used, the mLDP procedures are exactly the same as described in [mLDP], with the following exception. If LSR D is setting up MP-LSP <R,X>, its "upstream LSR" is selected according to the procedures of section 1.3.1, and is not necessarily the "IGP next hop" on D's path to R.

Suppose that LSRs D1 and D2 are both setting up the P2MP MP-LSP

<R,X>, and that LSR U is the upstream LSR on each of their paths to R. D1 and D2 each binds a label to <R,X>, and each uses a label mapping message to inform U of the label binding. Suppose D1 has assigned label L1 to <R,X> and D2 has assigned label L2 to <R,X>. (Note that L1 and L2 could have the same value or different values; D1 and D2 do not coordinate their label assignments.) When U has a packet to transmit on the MP-LSP <R,X>, it makes a copy of the packet, pushes on label L1, and unicasts the resulting packet to D1. It also makes a second copy of the packet, pushes on label L2, and then unicasts the resulting packet to D2.

This procedure also works when the MP-LSP <R,X> is a MP2MP LSP. Suppose that in addition to labels L1 and L2 described above, U has assigned label L3 for <R,X> traffic received from D1, and label L4 for <R,X> traffic received from D2. When U processes a packet with label L3 at the top of its label stack, it knows the packet is from D1, so U sends a unicast copy of the packet to D2, after swapping L3 for L2. U does not send a copy back to D1.

Note that all labels used in this procedure are downstream-assigned labels.

The method of unicast is a local matter, outside the scope of this specification. The only requirement is that D1 will receive the copy of the packet carrying label L1, and that D1 will process the packet by looking up label L1. (And similarly, D2 must receive the copy of the packet carrying label L2, and must process the packet by looking up label L2.)

Note that if the method of unicast is MPLS, U will need to push another label on each copy of the packet before transmitting it. This label needs to ensure that delivery of the packet to the appropriate LSR, D1 or D2. Use of penultimate-hop popping for that label is perfectly legitimate.

3. Targeted mLDP with Multicast Tunneling

Suppose that LSRs D1 and D2 are both setting up MP-LSP <R,X>, and that LSR U is the upstream LSR on each of their paths to R. Since multicast tunneling is being used, when U has a packet to send on this MP-LSP, it does not necessarily send two copies, one to D1 and one to D2. It may send only one copy of the packet, which will get replicated somewhere downstream in the multicast tunnel. Therefore, the label that gets bound to the MP-LSP must be an upstream-assigned label, assigned by U. This requires a change from the procedures of [mLDP]. D1 and D2 do not send label mapping messages to U; instead they send label request messages to U, asking U to assign a label to

the MP-LSP <R,X>. U responds with a label mapping message containing an upstream-assigned label, L (using the procedures specified in [LDP-UP]). As part of the same label mapping message, U also sends an Interface TLV (as specified in [LDP-UP]) identifying the multicast tunnel in which data on the MP-LSP will be carried. When U transmits a packet on this tunnel, it first pushes on the upstream-assigned label L, and then pushes on the label that corresponds to the multicast tunnel.

If the numerical value L of the upstream-assigned label is the value 3, defined in [LDP] and [RFC3032] as "Implicit NULL", then the specified multicast tunnel will carry only the specified MP-LSP. That is, aggregation of multiple MP-LSPs into a single multicast tunnel is not being done. In this case, no upstream-assigned label is pushed onto a packet that is transmitted through the multicast tunnel.

Various types of multicast tunnel may be used. The choice of tunnel type is determined by provisioning, or by some other method that is outside the scope of this document. [LDP-UP] specifies encodings allowing U to identify an mLDP MP-LSP, and RSVP-TE P2MP LSP, as well as other types of multicast tunnel.

This document does not specify procedures for tunneling one or more MP2MP LSPs through P2MP tunnels. While it is possible to do this, it is highly RECOMMENDED that MP2MP LSPs be tunneled through MP2MP LSPs (unless, of course, unicast replication is being used).

If the multicast tunnel is an mLDP MP-LSP or an RSVP-TE P2MP LSP, when U transmits a packet on the MP-LSP <R,X>, the upstream-assigned label L will be the second label in the label stack. Penultimate-hop popping MUST NOT be done, because the top label provides the context in which the second label is to be interpreted. See [RFC5331].

When LSR U uses these procedures to inform LSR D that a particular MP-LSP is being carried in a particular multicast tunnel, U and D MUST take appropriate steps to ensure that packets U sends into this tunnel will be received by D. The exact steps to take depend on the tunnel type. As long as U is D's upstream LSR for any MP-LSP that has been assigned to this tunnel, D must remain joined to the tunnel.

Note that U MAY assign the same multicast tunnel for multiple different MP-LSPs. However, U MUST assign a distinct upstream-assigned label to each MP-LSP. This allows the packets traveling through the tunnel to be demultiplexed into the proper MP-LSPs.

If U has an MP-LSP <R1,X1> with downstream LSRs D1 and D2, and an MP-LSP <R2,X2> with downstream LSRs D2 and D3, U may assign both MP-LSPs

to the same multicast tunnel. In this case, D3 will receive packets traveling on <R1,X1>. However, the upstream-assigned label carried by those packets will not be recognized by D3, hence D3 will discard those packets. Similarly, D1 will discard the <R2,X2> packets.

This document does not specify any rules for deciding whether to aggregate two or more MP-LSPs into a single multicast tunnel. Such rules are outside the scope of this document.

Except for the procedures explicitly details in this document, the procedures of [mLDP] and [LDP-UP] apply unchanged.

4. IANA Considerations

This document has no considerations for IANA.

5. Security Considerations

This document raises no new security considerations beyond those discussed in [LDP], [LDP-UP], and [RFC5331].

6. Acknowledgments

The authors wish to think Lizhong Jin for his comments.

7. Authors' Addresses

Maria Napierala
AT&T Labs
200 Laurel Avenue, Middletown, NJ 07748
E-mail: mnapierala@att.com

Eric C. Rosen
Cisco Systems, Inc.
1414 Massachusetts Avenue
Boxborough, MA, 01719
E-mail: erosen@cisco.com

IJsbrand Wijnands
Cisco Systems, Inc.
De kleetlaan 6a Diegem 1831
Belgium
E-mail: ice@cisco.com

8. Normative References

[LDP] Loa Andersson, Ina Minei, Bob Thomas, editors, "LDP Specification", RFC 5036, October 2007

[LDP-CAP] Bob Thomas, Kamran Raza, Shivani Aggarwal, Rahul Aggarwal, Jean-Louis Le Roux, "LDP Capabilities", RFC 5561, July 2009

[mLDP] Ina Minei, Kireeti Kompella, IJsbrand Wijnands, Bob Thomas, "Label Distribution Protocol Extensions for Point-to-Multipoint and Multipoint-to-Multipoint Label Switched Paths", draft-ietf-mpls-ldp-p2mp-13.txt, April 2011

[LDP-UP] Rahul Aggarwal, Jean-Louis Le Roux, "MPLS Upstream Label Assignment for LDP", draft-ietf-mpls-ldp-upstream-10.txt, February 2011

[RFC2119] "Key words for use in RFCs to Indicate Requirement Levels.", Bradner, March 1997

[RFC3032] Eric Rosen, et. al., "MPLS Label Stack Encoding", RFC 3032, January 2001

[RFC5331] Rahul Aggarwal, Yakov Rekhter, Eric Rosen, "MPLS Upstream Label Assignment and Context-Specific Label Space", RFC 5331, August 2009

IETF
Internet Draft

Ping Pan
Rajan Rao
Biao Lu
(Infinera)
Luyuan Fang
(Cisco)
Andy Malis
(Verizon)
Sam Aldrin
(Huawei)
Mohana Singamsetty
(Tellabs)

Expires: January 11, 2012

July 11, 2011

Supporting Shared Mesh Protection in MPLS-TP Networks

draft-pan-shared-mesh-protection-02.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/lid-abstracts.txt>

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

This Internet-Draft will expire on January 11, 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.

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.....	3
2. Background.....	4
3. Problem Definition.....	5
4. Protection Switching.....	6
5. Activation Operation Overview.....	8
6. Protocol Definition.....	9
6.1. Activation Messages.....	9
6.2. Message Encapsulation.....	10
6.3. Reliable Messaging.....	11
6.4. Message Scoping.....	12
7. Processing Rules.....	12
7.1. Enable a protecting path.....	12
7.2. Disable a protecting path.....	13
7.3. Get protecting path status.....	14
7.4. Acknowledgement with STATUS.....	14
7.5. Preemption.....	14
8. Security Consideration.....	14
9. IANA Considerations.....	15
10. Normative References.....	15
11. Acknowledgments.....	15

1. Introduction

Shared mesh protection is a common traffic protection 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. The framework version of the document has been presented in ITU-T SG15 Interim Meeting in May 2011, and is in-sync with the on-going G.SMP work in ITU-T.

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. The activation messages may either traverse through a "trusted" transport channel, or require some level of built-in reliability 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.

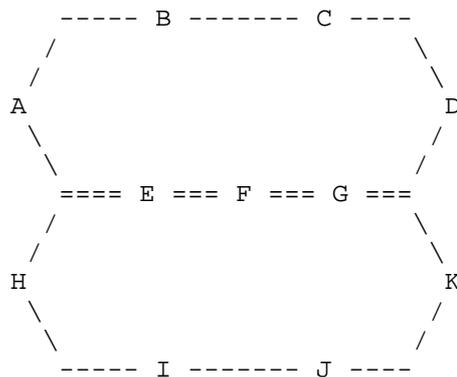
The activation signaling defined in this memo is to support warm standby in the context of MPLS-TP.

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'.

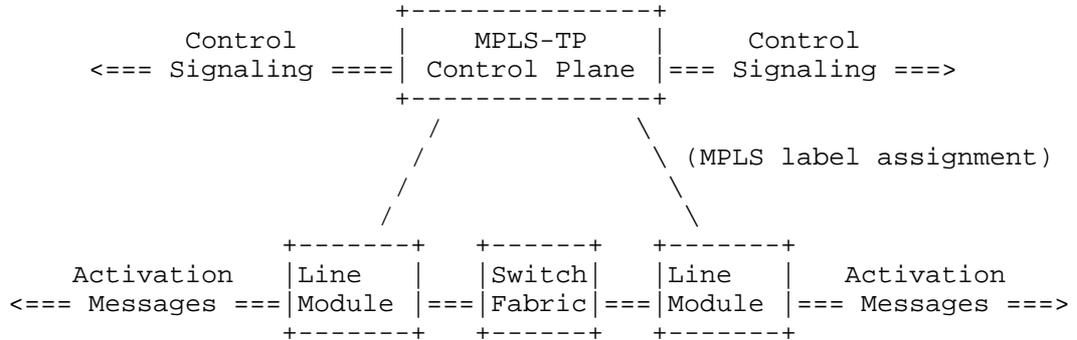
The operation may have extra caveat:

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.

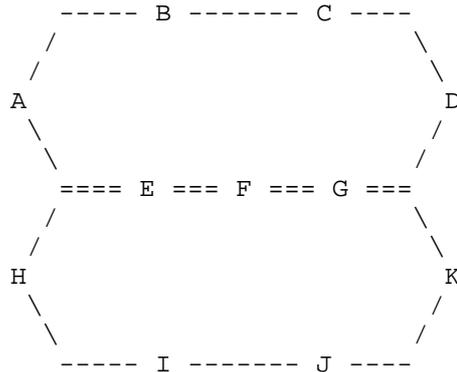
To achieve the tight end-to-end switch-over budget, it's possible to implement the entire activation procedure with hardware-assistance (such as in FPGA or ASIC).

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

To achieve high performance, the activation procedure is designed to be simple and straightforward on the network nodes.

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 may periodically poll the protecting paths about resource availability. The intermediate nodes have the option to inform the current resource utilization. This procedure may be conducted by other OAM mechanisms.

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.

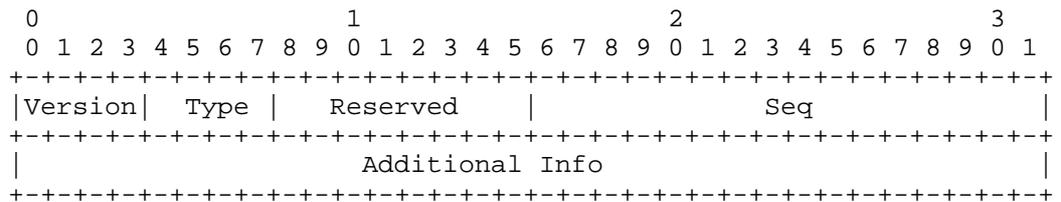
6. Protocol Definition

6.1. Activation Messages

The activation requires the following messages:

- o ENABLE: this is initiated by the headend nodes to activate a protecting path
- o DISABLE: this is initiated by the headend nodes to disable a protecting path and free the associated network resources
- o GET: this is initiated by the headend to gather resource availability information on a particular protecting path
- o NOTIFY: this is initiated by the intermediate nodes and terminate on the headend nodes to report preemption or protection failure conditions
- o 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:

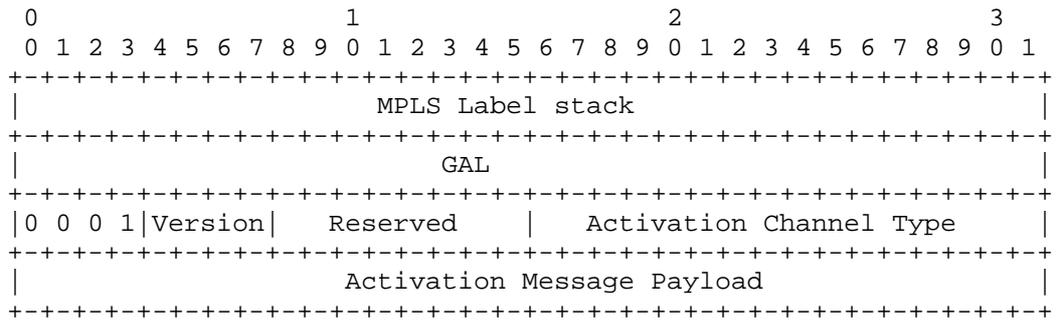


- o Version: 1
- o Type:

- o ENABLE 1
- o DISABLE 2
- o GET 3
- o STATUS 4
- o NOTIFY 5
- o Reserved: This field is reserved for future use
- o Seq: This uniquely identifies a particular message. This field is defined to support reliable message delivery
- o Additional Info: the message-specific data

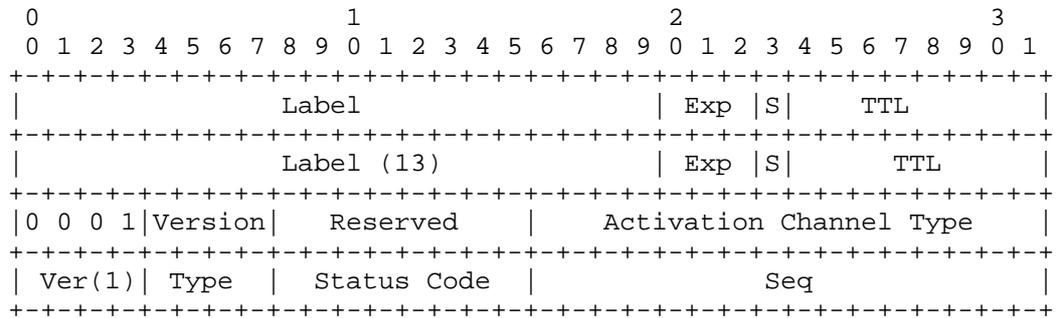
6.2. Message Encapsulation

Activation messages use MPLS labels to identify the paths. Further, the messages are encapsulated in GAL/GACH:



- o GAL is described in [3]
- o Activation Channel Type is the GACH channel number assigned to the protocol. This uniquely identifies the activation messages.

Specifically, the messages have the following message format:



For STATUS and NOTIFY messages, the Status Code has the following encoding value and definition:

- o 0-19: OK
 - . 1: end-to-end ack
- o 20-39: message processing errors
 - . 20: no such path
- o 40-59: processing issues:
 - . 40: no more resource for the path
 - . 41: preempted by another path
 - . 42: system failure
- o 60-79: informative data:
 - . 60: 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:

- o On receive, if the message has label TTL = 0, the node must drop the packet without further processing
- o 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)
- o 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 interval, 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.

During ITU-T SG15 Interim meeting in May 2011, we have had long discussion with the G.SMP contributors, in particular Fatai Zhang, Bin Lu, Maarten Vissers and Jeong-dong Ryoo. We thank their feedback and corrections.

Authors' Addresses

Ping Pan
Email: ppan@infinera.com

Rajan Rao
Email: rrao@infinera.com

Biao Lu
Email: blu@infinera.com

Luyuan Fang
Email: lufang@cisco.com

Andy Malis
Email: andrew.g.malis@verizon.com

Sam Aldrin
Email: sam.aldrin@huawei.com

Sri Mohana Satya Srinivas Singamsetty
Email: SriMohanS@Tellabs.com

MPLS Working Group
Internet Draft
Intended status: Informational
Expires: January 2012

Pranjal Kumar Dutta
Wim Henderickx
Alcatel-Lucent
July 4, 2011

Upstream LSR Redundancy for Multi-point LDP Tunnels
draft-pdutta-mpls-mldp-up-redundancy-00.txt

Status of this Memo

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

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

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

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

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

This Internet-Draft will expire on January 4, 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.

Abstract

Label Distribution Protocol (LDP) can be used to set up Point-to-Multipoint (P2MP) and Multipoint-to-Multipoint (MP2MP) Label Switched Paths. The existing specification for this functionality assumes that a downstream LSR selects only one upstream LSR for a P2MP or MP2MP LSP. A Make-Before-Break (MBB) procedure in mLdp base specification [MLDP] for graceful upstream LSR change but that is not applicable when the upstream LSR node fails. As IPTV deployments grow in number and size, service providers are looking for solutions that minimize the service disruption due to such failures. This document describes a set of procedures that minimize packet loss when an upstream LSR node fails. This document does not change any specifications of mLdp protocol as defined in [MLDP] and so there are no inter-operability requirements for the procedures described in this document.

Table of Contents

1. Introduction.....	2
1.1. Problem Statement.....	2
2. Conventions used in this document.....	3
3. Terminology.....	3
4. Upstream LSR Redundancy.....	4
5. Backup Upstream LSR Selection.....	5
5.1 Equal-Cost-Multi-Path (ECMP).....	5
5.2 Loop-Free-Alternate (LFA).....	6
6. Fast Failover to Backup Upstream LSR.....	7
7. Security Considerations.....	7
8. IANA Considerations.....	7
9. References.....	8
9.1. Normative References.....	8
9.2. Informative References.....	8
10. Acknowledgments.....	8

1. Introduction

1.1. Problem Statement

The Label Distribution Protocol (LDP) extensions for setting up Point-to-Multipoint (P2MP) Label Switched Paths (LSPs) and Multipoint-to-Multipoint (MP2MP) LSPs are specified in [mLdp]. This set of extensions is generally known as "Multipoint LDP" (mLdp) and this documents refer P2MP and MP2MP LSPs as "mLdp Tunnels", unless specified otherwise.

A node Z that wants to join an mLdp Tunnel determines the upstream peer U which is Z's next-hop on the best path from Z to the root node R of the mLdp tunnel. If there is more than one such LDP peer due to Equal-Cost-Multi-Path (ECMP), only one of them is picked. As defined in [MLDP] when there are several candidate upstream LSRs, the LSR Z must select only one upstream LSR based on a localized selection algorithm at node Z.

When the best path to reach the root changes, the mLdp tunnel may be broken temporarily resulting in packet loss until the LSR Z re-converges to a new upstream LSR U'. A set of Make-Before-Break (MBB) procedures is defined in [MLDP] for graceful transition to new upstream LSR U' and thus minimize this traffic loss. However these set of procedures are not applicable when upstream LSR U fails and that results in loss of traffic till topology re-converges to new upstream U' and Z completes set-up of new path towards the root. mLdp tunnels carry loss sensitive traffic such as broadcast video so it is very important to provide protection against such upstream node failures. A router's IGP convergence time is generally on the order of hundred's of milliseconds; the application traffic may be sensitive to losses greater than tens of milliseconds till the local LSR selects new upstream LSR and establishes the mLdp tunnel path towards the root node.

Minimizing traffic loss requires a mechanism for the local LSR Z on detection of failure to its immediate upstream U to rapidly invoke a repair path, which is minimally affected by any subsequent re-convergence. This document describes a set of procedures to provide protection against failure of upstream LSR by pre-establishing an mLdp tunnel path to one or more secondary upstream LSRs. The specification in this document leverage existing methodologies to achieve upstream LSR failover protection without imposing requirements on inter-operability or new protocol specification.

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

Primary Upstream : Upstream LSR chosen by a node to receive traffic on an mLdp tunnel.

Backup Upstream : Redundant Upstream LSR chosen by a node to receive traffic on an mLdp tunnel on failure of its primary upstream.

ECMP : Equal Cost Multi-Path

LFA : Loop-Free-Alternates

4. Upstream LSR Redundancy

Upstream LSR failure protection can be provided by taking advantage of redundant topologies in service provider networks. A local LSR Z selects two upstream LSRs - one primary LSR U and at least one backup LSR U'. Label mappings L sent to U and L' sent to U shares the same downstream next-hop label forwarding entries at Z. It needs to be ensured that mLdp tunnel path along such secondary LSR U' is loop free. Data packets are received by Z from both U and U' simultaneously. Redundant packets received from U' are discarded by Z. When Z detects a reachability failure to U then it switches its upstream to the backup LSR U' and packets are immediately available to forward out of each downstream next-hops. The amount of traffic loss in such failovers is dependent on the detection methodologies used by Z to detect failure of U. Details of such methodologies are out of scope of this document.

This mode of upstream LSR protection causes traffic replication from U' to Z which is dropped at Z. Note that such traffic replication does may demand extra capacity although on failure of U, the mLdp tunnel would have converged to U' anyway as a result of network topology change. In service provider networks it is likely that such redundant paths are kept disjoint from fate sharing. Redundant replications may not require additional capacity in the network, but may change replication distribution in upstream router U'.

This method is simple and is localize to individual routers along the path from root to leaves of an mLdp tunnels. There arises no interoperability requirements since there is no change in mLdp protocol operation. End-to-end failure protection and recovery can be achieved by such localized protection at every node.

this does not cause any loop in the path of the mLdp tunnel through U' from root node R towards the leaves.

5.2 Loop-Free-Alternate (LFA)

It is possible that in service provider network there are several loop free alternate paths exist. [RFC5286] provides specification for IP Fast Reroute with Loop-Free Alternates (LFA).

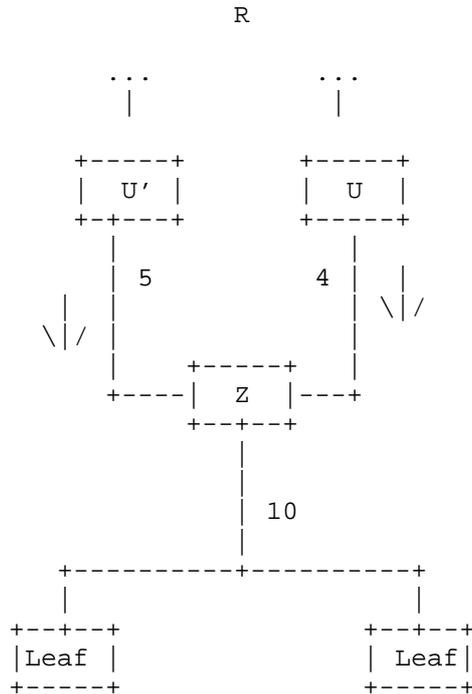


Figure 2.

In the absence of ECMP paths from Z towards root node R, IGPs may compute LFA paths from Z towards the root node. Z would select LDP peer on primary next-hop provided by IGP as primary upstream LSR U and peers on LFA next-hops as candidate backup upstream LSRs U'. Rapid upstream failure protection is achieved through use of pre-calculated backup next-hops that are loop-free and safe to use until the distributed network convergence process completes. This simple approach does not require any support from other routers. The extent to which this goal can be met by this specification is dependent on

the LFA topology of the network. Especially when networks does not have ECMP, then backup stream LSR selection using LFA has significant advantages. It is RECOMMENDED that LFA next-hops computed for this purpose are "Node-Protecting", that is the backup LSR U' in turn must not choose U as its primary upstream LSR. However without node protection although upstream node failure may not be protected, it definitely provides link protection on failure of downstream link from U to Z.

6. Fast Failover to Backup Upstream LSR

When node Z selects a backup upstream LSR and sends backup label mapping L' for joining the mLdp tunnel path towards the root, it is RECOMMENDED that Z installs L' into data plane with one exception : L' MUST NOT be forwarding traffic to its downstream - it is kept in "blocking mode". When Z detects failure on its primary upstream U, it triggers switchover of traffic from primary upstream label L to L', thus blocking L and unblocking L' in forwarding traffic to downstream(s). This is important to avoid duplication of traffic to downstream during failover. A single failover trigger may be sufficient for fast switchover of traffic in all mLdp tunnels that have selected U as primary upstream to their respective backup upstream labels.

For some implementations it may not be possible to pre-install a backup label L' into data plane in blocking mode. On primary upstream failure, if L' is added before L is removed, there is a potential risk of packet duplication, and/or the creation of transient data plane forwarding loop. If L is removed before L' is added, packet loss may result. For such implementations the RECOMMENDED procedure is to remove L before adding L'.

7. Security Considerations

This document does not require additional security considerations to what is required in [MLDP].

8. IANA Considerations

There is no IANA considerations required by this document.

9. References

9.1. Normative References

[MLDP] Ina Minei, Kireeti Kompella, IJsbrand Wijnands, Bob Thomas, "Label Distribution Protocol Extensions for Point-to-Multipoint and Multipoint-to-Multipoint Label Switched Paths", draft-ietf-mpls-ldp-p2mp-11.txt, October 2010

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

[RFC5286] A. Atlas, A. Zinin, "Basic Specification for IP Fast Re Route: Loop-Free Alternates.

9.2. Informative References

TBD.

10. Acknowledgments

Authors would like to acknowledge reviews and valuable feedbacks from Paul Kwok.

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

Authors' Addresses

Pranjal Kumar Dutta
Alcatel-Lucent
701, E Middlefield Road,
Mountain View, CA 94043,
USA.

Email: pranjal.dutta@alcatel-lucent.com

Wim Henderickx
Alcatel-Lucent
Copernicuslaan 50
2018 Antwerp, Belgium

Email: wim.henderickx@alcatel-lucent.be

Network Working Group
Internet Draft
Updates: 5036, 4447 (if approved)
Intended status: Standards Track
Expires: January 10, 2012

Kamran Raza
Sami Boutros
Luca Martini
Cisco Systems, Inc.

Nicolai Leymann
Deutsche Telekom

July 11, 2011

Applicability of LDP Label Advertisement Mode

draft-raza-mpls-ldp-applicability-label-adv-01.txt

Abstract

An LDP speaker negotiates the label advertisement mode with its LDP peer at the time of session establishment. Although different applications sharing the same LDP session may need different modes of label distribution and advertisement, there is only one type of label advertisement mode that is negotiated and used per LDP session. This document clarifies the use and the applicability of session's negotiated label advertisement mode, and categorizes LDP applications into two broad categories of negotiated mode-bound and mode-independent applications. This document proposal and clarification thus updates [RFC5036] and [RFC4447].

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/lid-abstracts.txt>

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

This Internet-Draft will expire on January 10, 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	3
2. Conventions used in this document	3
3. Label Advertisement Mode Applicability	4
3.1. Label Advertisement Mode Negotiation	4
3.2. LDP Applications Categorization	4
3.2.1. Session mode-bound Applications	5
3.2.2. Session mode-independent Applications	5
3.3. Update to RFC-5036	6
3.4. Update to RFC-4447	6
4. Future Work	6
5. Security Considerations	6
6. IANA Considerations	7
7. References	7
7.1. Normative References	7
7.2. Informative References	7
8. Acknowledgments	7

1. Introduction

The MPLS architecture [RFC3031] defines two modes of label advertisement for an LSR:

1. Downstream-on-Demand
2. Unsolicited Downstream

The "Downstream-on-Demand" mode requires an LSR to explicitly request the label binding for FECs from its peer, whereas "Unsolicited Downstream" mode allows an LSR to distribute the label binding for FECs unsolicitedly to LSR peers that have not explicitly requested them. The MPLS architecture [RFC3031] also specifies that on any given label distribution adjacency, the upstream LSR and the downstream LSR must agree to using a single label advertisement mode.

Label Distribution Protocol (LDP) [RFC5036] allows label advertisement mode negotiation at the session establishment time (section 3.5.3 [RFC5036]). To comply with MPLS architecture, LDP specification also dictates that only one label advertisement mode is agreed and used on a given LDP session between two LSRs.

With the advent of new applications, such as L2VPN [RFC4447], mLDP [MLDP], ICCP [ICCP], running on top of LDP, there are situations when an LDP session is shared across more than one application to exchange label bindings for different type of FECs. Although different applications sharing the same LDP session may need different type of label advertisement mode negotiated, there is only one type of label advertisement mode that is negotiated and agreed at the time of establishment of LDP session.

This document clarifies the use and the applicability of session's label advertisement mode for each application using the session. It also categorizes LDP applications into two broad categories of negotiated mode-bound and mode-independent applications. This document proposal and clarification thus updates [RFC5036] and [RFC4447].

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 unqualified term "mode" used in document refers to "label advertisement mode".

Please also note that LDP specification [RFC5036] uses the term "Downstream Unsolicited" to refer to "Unsolicited Downstream", as well as uses the terms "label distribution" and "label advertisement" interchangeably. This document also uses these terms interchangeably.

3. Label Advertisement Mode Applicability

3.1. Label Advertisement Mode Negotiation

Label advertisement mode is negotiated between participating LSR peers at the time of session establishment. The label advertisement mode is specified in LDP Initialization message's "Common Session Parameter" TLV by setting A-bit (Label Advertisement Discipline bit) to 1 or 0 for Downstream-on-Demand or Downstream-Unsolicited modes respectively [RFC5036]. The negotiation of the A-bit is specified in section 3.5.3 of [RFC5036] as follows:

"If one LSR proposes Downstream Unsolicited and the other proposes Downstream on Demand, the rules for resolving this difference is:

- If the session is for a label-controlled ATM link or a label-controlled Frame Relay link, then Downstream on Demand MUST be used.

- Otherwise, Downstream Unsolicited MUST be used."

Once label advertisement mode has been negotiated and agreed, both LSRs must use the same mode for label binding exchange.

3.2. LDP Applications Categorization

At the time of standardization of LDP base specification RFC-3036, the earlier applications using LDP to exchange their FEC bindings were:

- . Dynamic Label Switching for IP Prefixes
- . Label-controlled ATM/FR

Since then, several new applications have emerged that use LDP to signal their FEC bindings and/or application data:

- . L2VPN P2P PW ([RFC4447])

- . L2VPN P2MP PW ([P2MP-PW])
- . mLDP ([MLDP])
- . ICCP ([ICCP])

We divide these LDP applications into two broad categories from label advertisement mode usage point of view:

1. Session mode-bound Applications (i.e. use the negotiated label advertisement mode)
2. Session mode-independent Applications (i.e. do not care about the negotiated label advertisement mode)

3.2.1. Session mode-bound Applications

The FEC label binding exchange for such LDP applications MUST use the negotiated label advertisement mode.

The early LDP applications "Dynamic Label Switching for IP Prefixes" and "Label-controlled ATM/FR" fall into this category.

3.2.2. Session mode-independent Applications

The FEC label binding, or any other application data, exchange for such LDP applications does not care about, nor tied to the negotiated label advertisement mode of the session; rather, the information exchange is driven by the application need and procedures as described by their respective specifications. For example, [MLDP] specifies procedures to advertise P2MP FEC label binding in an unsolicited manner, irrespective of the negotiated label advertisement mode of the session.

The applications, PW (P2P and P2MP), MLDP, and ICCP, fall into this category of LDP application.

3.2.2.1. Upstream Label Assignment

As opposed to downstream assigned label advertisement defined by [RFC3031], [LDP-UPSTREAM] specification defines new mode of label advertisement where label advertisement and distribution occurs for upstream assigned labels.

As stated in earlier section 3.1 of this document, [RFC5036] only allows specifying Downstream-Unsolicited or Downstream-on-Demand mode. This means that any LDP application that requires upstream

assigned label advertisement also falls under the category of Session mode-independent application.

3.3. Update to RFC-5036

For clarification reasons, section 3.5.3 of [RFC5036] is updated to add following two statements under the description of "A, Label Advertisement Discipline":

- The negotiated label advertisement discipline only applies to FEC label binding advertisement of "Address Prefix" FECs;
- Any document specifying a new FEC SHOULD state the applicability of the negotiated label advertisement discipline for that FEC.

3.4. Update to RFC-4447

[RFC4447] specifies LDP extensions and procedures to exchange label bindings for P2P PW FECs. The section 3 of [RFC4447] states:

"LDP MUST be used in its downstream unsolicited mode."

Since PW application falls under session mode-independent application category, the above statement in [RFC4447] should be read to mean as follows:

"LDP MUST exchange PW FEC label bindings in downstream unsolicited manner, independent of the negotiated label advertisement mode of the LDP session."

4. Future Work

This document only clarifies the existing behavior for LDP label advertisement mode for different applications without defining any protocol extensions. In future, a new LDP capability [RFC5561] based mechanism can be defined to signal/negotiate label advertisement mode per FEC/application.

5. Security Considerations

This document specification only clarifies the applicability of LDP session's label advertisement mode, and hence does not add any LDP security mechanics and considerations to those already defined in LDP specification [RFC5036].

6. IANA Considerations

None.

7. References

7.1. Normative References

- [RFC5036] Andersson, L., Minei, I., and Thomas, B., Editors, "LDP Specification", RFC 5036, September 2007.
- [RFC3031] Rosen, E., Viswanathan, A., and Callon, R., "Multiprotocol Label Switching Architecture", RFC 3031, January 2001.

7.2. Informative References

- [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.
- [P2MP-PW] Boutros, S., Martini, L., Sivabalan, S., Del Vecchio, G., Kamite, Jin, L., "Signaling Root-Initiated P2MP PWs using LDP", draft-ietf-pwe3-p2mp-pw-02.txt, Work in Progress, March 2011.
- [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-mppls-ldp-p2mp-14.txt, Work in Progress, June 2011.
- [ICCP] Martini, L., Salam, S., Sajassi, A., and Matsushima, S., "Inter-Chassis Communication Protocol for L2VPN PE Redundancy", draft-ietf-pwe3-iccp-05.txt, Work in Progress, April 2011.
- [UPSTREAM-LDP] Aggarwal, R., and Le Roux, J.L., "MPLS Upstream Label Assignment for LDP", draft-ietf-mppls-ldp-upstream-10.txt, Work in Progress, February 2011.
- [RFC5561] Thomas, B., Raza, K., Aggarwal, S., Aggarwal, R., and Le Roux, J.L., "LDP Capabilities", RFC 5561, July 2009.

8. Acknowledgments

The authors would like to acknowledge Eric Rosen and Rajiv Asati for their review and input.

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

Luca Martini
Cisco Systems, Inc.
9155 East Nichols Avenue, Suite 400,
Englewood, CO 80112, USA.
E-mail: lmartini@cisco.com

Nicolai Leymann
Deutsche Telekom,
Email: N.Leymann@telekom.de

Network Working Group
Internet Draft
Intended status: Standards Track
Expires: December 2, 2011

Kamran Raza
Sami Boutros
Pradosh Mohapatra

Cisco Systems, Inc.

June 3, 2011

LDP Outbound Label Filtering

draft-raza-mpls-ldp-olf-00.txt

Abstract

The Label Distribution Protocol (LDP) allows one Label Switching Router (LSR) to advertise to another a set of "bindings" between MPLS labels and "Forwarding Equivalence Classes" (FECs). Suppose LSR2 is advertising a set of label bindings to LSR1. Frequently, LSR1 does not need to know all of LSR2's label bindings, and LSR1 may be configured to disregard bindings in which it has no interest. This document defines an "Outbound Label Filtering" (OLF) mechanism that allows LSR1 to inform LSR2 dynamically of the set of FECs for which it needs to receive label bindings. LSR2 then applies this filter before sending its label bindings to LSR1. In addition to the generic aspects of this mechanism, this document also specifies outbound label filter for the "Address Prefix FEC" type.

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/lid-abstracts.txt>

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

This Internet-Draft will expire on December 2, 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.

Table of Contents

1. Introduction	3
2. Conventions used in this document	3
3. FEC Label Bindings	3
4. Outbound Label Filter	4
4.1. Constructs	4
4.1.1. FEC-Type	4
4.1.2. OLF Policy	5
4.2. OLF Signaling	6
4.2.1. OLF Policy Status TLV	6
4.2.2. OLF Element Format	7
4.2.3. OLF Entry Format	8
4.3. OLF Capability negotiation	10
4.4. OLF Procedures	12
4.4.1. OLF Capability Negotiation At Session Estab. Time	12
4.4.2. OLF Capability Dynamic Changes	13
4.4.3. OLF Policy Updates	15
5. Address Prefix FEC OLF Type	16
5.1. Matching Address Prefixes to OLF Entries	17
6. Operational Examples	18
6.1. Label Filtering at Area Border Router	18
6.2. LSR with limited LIB size	19
7. Security Considerations	19
8. IANA Considerations	19
9. References	20
9.1. Normative References	20
9.2. Informative References	20
10. Acknowledgments	20

1. Introduction

The Label Distribution Protocol (LDP) allows one Label Switching Router (LSR) to advertise to another a set of "bindings" between MPLS labels and "Forwarding Equivalence Classes" (FECs). When LDP's "Downstream Unsolicited" mode [RFC5036] is in use, an LSR may receive label bindings for FECs in which it has no interest. The receiving LSR typically filters out these unwanted label bindings based on its local policy. Since the advertisement of label binding updates by the sender, as well as the processing of these updates by the receiver, consume network bandwidth and LSR resources, it may be beneficial if the advertisement of such label bindings can be avoided at the source itself under the control of the receiver.

This document defines a label filtering mechanism that allows an LDP speaker to send to its LDP peer a set of FEC-based Outbound Label Filters (OLFs). The peer would apply these filters, in addition to any local outbound filtering policy, to constrain/filter its outbound label binding updates to the speaker.

This document also defines the Outbound Label Filter (OLF) type "Address Prefix FEC Outbound Label Filter" for "Address Prefix FEC" type, which can be used to perform label filtering for IP Prefix label bindings.

This specification is modeled on [RFC5291] and [RFC5292].

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 "FEC-Type" is used to refer to a tuple consisting of <FEC Element Type, Address Family>.

3. FEC Label Bindings

MPLS LDP associates a FEC with each Label Switched Path (LSP) it creates [RFC5036]. This means that a label is assigned for 1 or more FEC(s) and label bindings advertised to peers are bound to FEC(s). To define an LDP OLF, filters need to be defined for label bindings.

These filter definitions need to include both FEC Element type, as well as address family, if/as applicable, for a given FEC type.

Following is a list of most commonly used LDP FEC elements at the time of writing of this document:

FEC Element Type	Address Family	Specification
-----	-----	-----
Wildcard	N/A	[RFC5036]
Address Prefix	IPv4, IPv6	[RFC5036]
Typed Wildcard	AF of Sub-FEC	[RFC5918]
P2MP	IPv4, IPv6	[mLDP]
MP2MP-Upstream	IPv4, IPv6	[mLDP]
MP2MP-Downstream	IPv4, IPv6	[mLDP]
PWid	N/A	[RFC4447]
Generalized PWid	N/A	[RFC4447]
P2MP PW	N/A	[P2MP-PW]

Table 1: LDP FEC Types

This document defines a framework for label filtering that applies to all of the FEC types listed under Table 1, except "Wildcard" and "Typed Wildcard" FEC types. The framework is also easily extensible for new FEC types that may get defined in the future.

4. Outbound Label Filter

4.1. Constructs

4.1.1. FEC-Type

In the context of this document, we define "FEC-Type" as a construct that uniquely identifies (or maps to) a FEC. This is defined as a tuple of the following form:

```
<FEC Element Type, Address Family>
```

As shown in Table 1, not all FEC elements require qualification with Address Family. For those types, the address family is not specified (set to a reserved value).

Following are some example of FEC-Types:

```
<Address Prefix FEC Element, IPv4>
```

<Address Prefix FEC Element, IPv6>

<Pwid, N/A>

4.1.2. OLF Policy

We define an Outbound Label Filtering (OLF) Policy as a set of one or more OLF Elements each corresponding to a given FEC-Type. Where, an OLF Element itself comprises one or more OLF Entries.

4.1.2.1. OLF Element

An OLF Element is identified by a FEC-Type and consists of one or more OLF entries that have a common FEC-Type. The "FEC-Type" component uniquely identifies a FEC and is used to provide a coarse granularity control by limiting an OLF to only those FECs that match the FEC-Type component.

To define an OLF Element for a given FEC-Type, precise conditions and rules need to be specified under which the given FEC is considered to match a particular OLF entry.

4.1.2.2. OLF Entry

An OLF entry is a tuple of the form:

<Action, OLF-value>

The "Action" component specifies how the OLF filter is to be handled by the receiving LSR. The specified values for Action include "PERMIT", "DENY", and "PERMIT-ALL". PERMIT action indicates to receiving LSR to allow advertisement of label bindings for the set of FECs that match the OLF entry, DENY is opposite of PERMIT and disallows (i.e. filters) the advertisement of label bindings for the set of FECs that match the OLF entry. PERMIT-ALL is the wildcard equivalent of PERMIT, and hence apply to all FECs associated with the FEC-Type of the OLF Element corresponding to OLF entry.

The "OLF-value" component is FEC-specific and provides the specification of FEC for matching. This component is not mandatory and is not present when Action component is PERMIT-ALL. The format of OLF-value for a FEC element type is to be defined by the designer of the given FEC element. This document defines the format of OLF-Value for FEC-Types corresponding to "Address Prefix" FEC Element type [RFC5036].

4.2. OLF Signaling

4.2.1. OLF Policy Status TLV

An OLF is signaled to a peer through LDP Notification messages. A new status TLV, named "OLF Policy Status", is introduced to carry the OLF specifications. This TLV is carried in the optional parameter section of the LDP Notification message. Moreover, a new LDP Status Code, "OLF Status", is defined for use in LDP Status TLV to indicate the presence of "OLF Policy Status" TLV in a given Notification message.

A single OLF Policy Status TLV may contain one or more OLF Element sub-TLVs. Each OLF Element TLV represents a single FEC-Type and consists of one or more "OLF Entry" sub-TLVs.

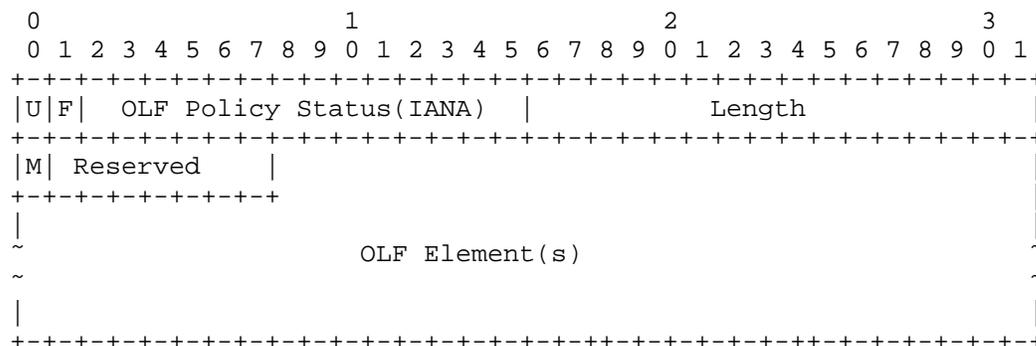


Figure 1: OLF Policy Status TLV

Where:

U/F bits: U-bit/F-bit MUST be set to 1/0 respectively so that a receiver MUST silently ignore this TLV if unknown to it, and continue processing the rest of the message.

Length: Total length (in octets) of "OLF Policy Status TLV" following the "Length" field. There is no padding requirement at the end of this TLV in case TLV does not end at Word boundary.

OLF Element(s): One or more OLF Element sub-TLVs. In a given OLF Policy Status TLV, only one OLF Element for a given FEC-Type is allowed. If more than one OLF Element is present for a given FEC-Type, then receiving LSR MUST pick the first occurrence of

such an element and ignore the other occurrences corresponding to the given FEC-Type.

M-bit: "More" bit specifying if there are more/further OLF Policy Status to follow for the given update set. The bit is set to 1 if there are further portion of policy that will follow in subsequent message(s), and set to 0 if the TLV alone constitutes the policy, or is the last update for the given update set.

Reserved bits: Reserved for future use. MUST be set to zero on transmit and MUST be ignored on receipt.

An LSR MAY also update its OLF with a peer by sending subsequent "OLF Policy Status" TLVs in LDP Notification messages. The receipt of an OLF Policy update from a peer for a given FEC-Type is meant to replace (overwrite) the previously installed FEC-Type OLF policy corresponding to the peer, if any, at the receiving LSR.

A complete OLF policy can be splitted across more than one OLF policy updates -- e.g. if the given OLF policy is big enough to fit in a single Notification message (due to LDP PDU size limitation [RFC5036]). In such cases, the sender LSR sends more than one LDP Notification message(s) with "OLF Policy Status" TLV, splitting the policy on OLF Element boundaries (i.e. an OLF Element MUST NOT span across more than one message). The sender also indicates if more than single Policy message will be sent for the given OLF update, as well as indicates the last message in the given update set. The receiver LSR, upon receiving OLF updates that span across more than one message, stores them in the order of receipt and processes them only after complete policy set has been received. If an LSR receives an incomplete/partial update set, and does not receive end of update (i.e. last message in the given set with M bit set to 0), it keeps these partial updates in its temporary buffer until one of the following events occur:

1. End of [policy] update received (OLF Policy Status TLV with M=0)
2. Session terminates
3. OLF capability changes

4.2.2. OLF Element Format

As shown in Figure 2, an OLF Element comprises one or more OLF entries grouped by FEC-Type <FEC Element Type, Address Family>:

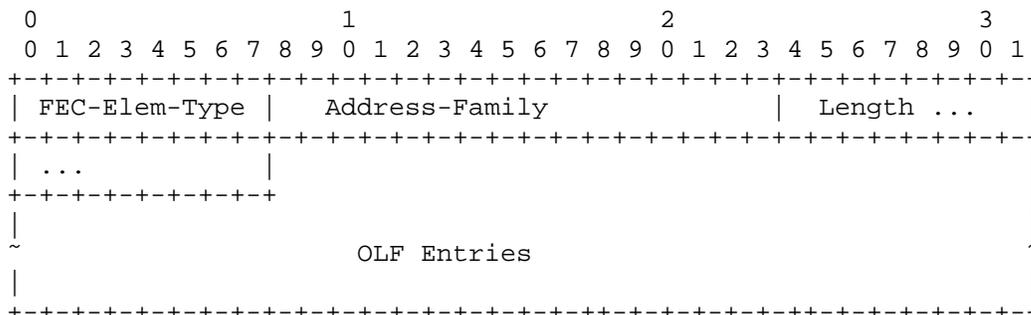


Figure 2: OLF Element format

Where:

FEC-Elem-Type/Address-Family: These fields jointly represent a FEC-Type. For the FEC element types listed in Table 1 which do not require Address Family qualification, Address-Family field MUST be set to zero on transmit and MUST be ignored on receipt.

Length: Length (in octets) of the OLF Element sub-TLV following the "Length" field; i.e. total length of OLF entries that follow in the given OLF Element sub-TLV. There is no padding requirement at the end of this TLV in case TLV does not end at Word boundary.

4.2.3. OLF Entry Format

Each OLF Entry is encoded as follows:

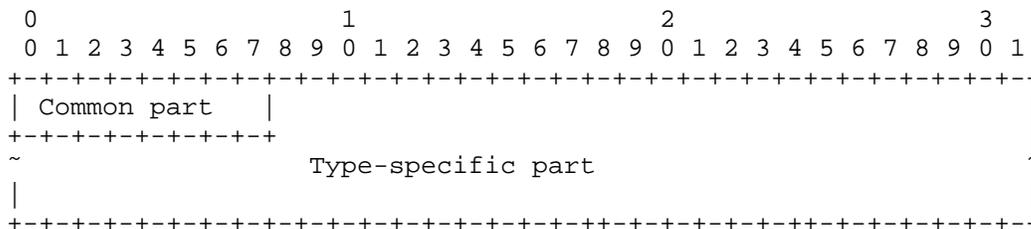


Figure 3: OLF Entry format

Where:

Common part: Common definition that is applicable to all types of OLF entries.

Type-specific part: Type specific (variable) definition corresponding to FEC-Type; also called "OLF-value" under section 4.1.2.2.

The "Common part" is one-octet field defined as following:

```

0 1 2 3 4 5 6 7
+---+---+---+---+
|Action | Rsvd  |
+---+---+---+---+

```

Where:

Action: Indicates the desired action (operation) to be performed by receiving LSR on received OLF entry, if FEC matches. The possible values are

```

0: PERMIT
1: DENY
2: PERMIT-ALL
4-15: Reserved (for future use).

```

Rsvd: Reserved for future use. MUST be set to 0 on transmit and MUST be ignored on the receipt.

4.2.4. Rules for OLF Element and OLF Entry

Following rules apply to OLF Element and Entries:

- o When the Action component of an OLF entry specifies a wildcard operation (PERMIT-ALL), then the OLF entry MUST consist of only the Common part.
- o When an OLF Element contains more than one OLF entry, then receiving LSR MUST process the OLF entries in the same order as they are specified inside the OLF element.
- o When processing a received OLF Element, an LSR MUST assume an implicit "DENY-ALL" as the last rule/entry. This assumption means that LSR denies all those FECs [of given FEC-Type] that have not already been matched in any of the specified OLF entries. This also means that the sender LSR needs to construct an OLF Element while keeping in mind an implicit DENY-ALL as the last rule.

4.3. OLF Capability negotiation

When a session has been negotiated to operate in Downstream Unsolicited mode, LDP speakers exchange all of their label bindings. If it is desired/required to exchange only selected label bindings between peers, the "Outbound Label Filtering Capability" is negotiated at session establishment time or at a later time.

An LDP speaker advertises the OLF Capability to announce to its peer its capability [and desire] to either send or receive or both send/receive OLF filters. The OLF feature will, however, work only when at least one LSR is able to send and other able to receive the OLF filters. The OLF Capability can be sent either in an Initialization message (Capability TLV's S-bit MUST be set to 1) or in a Capability message (Capability TLV's S-bit set to 1 or 0 to advertise or withdraw this capability respectively).

"Outbound Label Filtering" (OLF) capability is a new LDP capability, defined in accordance with LDP Capability definition guidelines [RFC5561]. The format of "Outbound Label Filtering" capability is as follows:

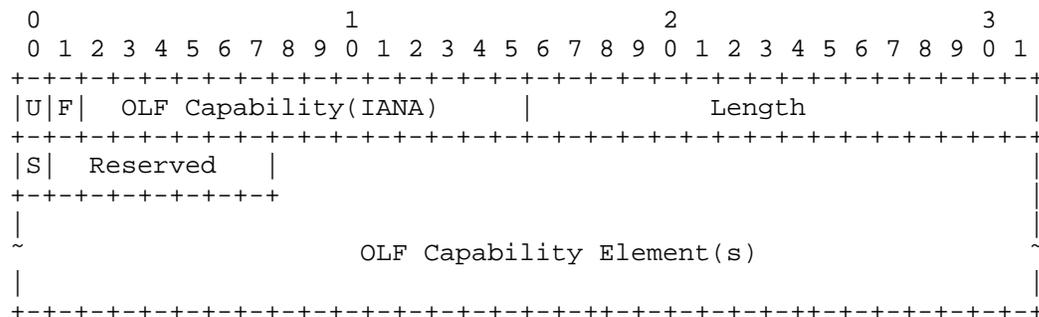


Figure 5: OLF Capability TLV

Where:

U/F-bits: The U-bit/F-bit for the TLV MUST be set to 1/0 respectively so that a receiver MUST silently ignore this TLV if unknown to it, and continue processing the rest of the message.

Length: The length (in octets) of TLV following "Length" field. The value of this field is variable because it depends on Capability-specific data [RFC5561] that follows in the TLV.

There is no padding requirement at the end of this TLV in case TLV does not end at Word boundary.

S-bit: The value of S-bit [RFC5561] is set to 1 or 0 to advertise or withdraw the capability respectively.

OLF Capability Element(s): This is the Capability-specific data [RFC5561] that is defined for OLF Capability, and consists of one or more "OLF Capability Element" types (defined below).

An LDP speaker that advertises OLF capability MUST support "OLF Policy Status" and "OLF Status" Status Code.

The format of an "OLF Capability Element" sub-TLV is specified as follows:

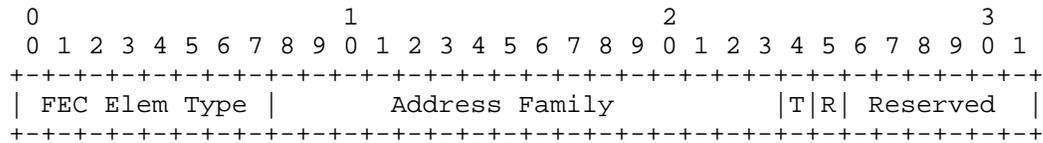


Figure 6: OLF Capability Element

Where:

FEC Elem Type / Address Family: These fields jointly represent a FEC-Type. For the FEC element types listed in Table 1 which do not require Address Family qualification, Address-Family field MUST be set to zero on transmit and MUST be ignored on receipt.

T-bit: Transmit/Send capability; set to 1 when the LDP speaker is able/willing to send OLF filters to its peer, set to zero otherwise.

R-bit: Receive capability; set to 1 when the LDP speaker is able/willing to receive OLF filters from its peer, set to zero otherwise.

Reserved: 6-bits reserved for future use. MUST be set to zero on transmit and MUST be ignored on receipt.

An LDP speaker SHOULD NOT send an "OLF Capability Element" with both T/R bits set to zero. If an LSR receives an OLF Capability Element

with both T/R bits set to zero, then the receiving LSR SHOULD ignore the corresponding OLF Capability Element and continue processing the rest of the TLV. The semantics and usage of T/R-bits is elaborated more in following sections.

There MUST be one and only one OLF Capability Element specified for a given FEC-Type in an OLF Capability TLV. Upon receiving more than one OLF Capability Element for a given FEC-Type in the same "OLF Capability TLV", the receiving LSR MUST send an LDP Notification message towards the sender with "Malformed TLV" status code, and abort the processing of entire message.

4.4. OLF Procedures

To describe the OLF procedures in the following subsections, let us consider LDP speaker LSR1 that is capable of sending OLF policy filters (for one or more FEC types), and LSR2 that is capable of receiving (and processing) them. Let us assume that the supported FEC-Types for OLF are IPv4/IPv6 "Address Prefix FEC" OLF types. Henceforth, both LSRs are configured respectively to send/receive OLF filters for "IPv4/IPv6 Address Prefix" OLF types to/from its peer. Let us also assume that the LSR1 is configured with an OLF filtering policy for "IPv4/IPv6 Address Prefix" FEC-Types that needs to be pushed to LSR2.

Moreover, assume that both LSR1 and LSR2 support "Dynamic Capability Announcement" capability TLV [RFC5561] and hence are capable of handling dynamic capability changes.

4.4.1. OLF Capability Negotiation At Session Establishment Time

At the session initialization time, LSR1 constructs an "OLF Capability TLV" with S-bit set to 1. The TLV also contains two OLF Capability Elements corresponding to FEC-Types "IPv4 Address Prefix" (FEC Elem Type=0x2, Address Family=0x1) and "IPv6 Address Prefix" (FEC Elem Type=0x2, Address Family=0x2). The LSR also sets T-bit/R-bit of these OLF Capability Elements to 1/0 respectively.

LSR1 then includes this "OLF Capability TLV" in the LDP Initialization message to LSR2.

LSR2, on the other hand, constructs/sends the "OLF Capability TLV" in the same manner as done by LSR1; the only difference being that LSR2 sets T-bit/R-bit of its OLF Capability Elements to 0/1 respectively.

Having exchanged/negotiated the "OLF Capability TLVs" successfully, LSR2 treats this as an implicit DENY for all label bindings for given FEC-Types (IPv4/IPv6 Prefix) and blocks any label binding advertisements towards LSR1 corresponding to these FEC-Types. LSR2 now waits for subsequent OLF filters/policy (via LDP Notification messages) from LSR1. LSR1 also understands that LSR2 is capable of receiving the OLF filters and hence it constructs OLF filters using its configured OLF policy for LSR2, and sends these filters to LSR2 via "OLF Policy Status TLV" in an LDP Notification message (Status code set to OLF Status). Upon the receipt of such an OLF policy, LSR2 reacts and applies the received outbound policy in addition to any locally configured outbound policy, and advertises towards LSR1 the label bindings corresponding to the matching "permitted" prefixes.

Since LSR2 is operating only in Receive mode for given OLF with LSR1, LSR1 does not block the advertisements and advertises all its label bindings for given IP Prefix FECs (in accordance with its locally configured outbound policy) towards LSR2.

4.4.1.1. Peer Incapable of "Receive" OLF

Consider a case where LSR2 is not OLF "Receive" capable for given FEC-Types. This means that LSR2 either does not send any "OLF Capability" corresponding to given FEC-Type, or "OLF Capability" for given FEC-Type does not have R-bit set. Having negotiated the "OLF Capability" for given FEC-Types, LSR1 realizes that LSR2 is not capable of receiving OLF filters for given FEC-Type(s), and hence LSR1 does not send any OLF filters (via LDP Notification message). In this case, LSR2 sends label bindings corresponding to given FEC-Type(s) towards LSR1 in unsolicited manner after session establishment, at which point, LSR1 may chose to discard them by applying the filtering policy in inbound direction.

4.4.2. OLF Capability Dynamic Changes

It is possible that OLF capability is enabled on an LSR after session has already been established with the peer. To signal and negotiate OLF Capability dynamically, both peers MUST support "Dynamic Capability Announcement" TLV [RFC5561].

4.4.2.1. "Send" OLF capability changes

Let's consider a case when LSR2 is initially configured to be able to receive OLF filters for IPv4/IPv6 Prefix FEC-Types, but LSR1 is not configured to be able to "send" the same. Now, a user enables and configures LSR1 to send OLF filters for given FECs towards LSR2.

This triggers LSR1 to construct an "OLF Capability" TLV in the same manner as described in section 4.4.1. The constructed "OLF Capability" is sent in a Capability message (with S-bit set to 1) towards LSR2. Upon receipt of this Capability message, LSR2 withdraws all label bindings from LSR1 corresponding to given FEC-Type(s). Later on, LSR1 sends its OLF filters via "OLF Policy Status" and duly applied by LSR2.

Assuming both LSR1 and LSR2 are already engaged in OLF filtering in sender and receiver roles respectively for given FEC-Types. Now consider that LSR1 configuration is changed to remove "send" capability for one FEC type (say IPv4 Prefix) towards LSR2. This triggers LSR1 to construct an "OLF Capability" TLV that includes only one OLF Capability Element corresponding to "IPv4 Prefix" FEC type. The constructed "OLF Capability" is sent in a Capability message (with S-bit set to 0) towards LSR2. Upon receipt of this Capability [withdrawal] message, LSR2 removes any existing OLF filter towards LSR1 corresponding to given FEC-Type "IPv4 Prefix", and re-advertises to LSR1 its entire label bindings database for given FEC-Type.

4.4.2.2. "Receive" OLF capability changes

Let's consider a case when LSR1 is initially configured to be able to send OLF filters for IPv4/IPv6 Prefix FEC-Types, but LSR2 is not configured to be able to "receive" the same. Now, a user enables and configures LSR2 to be able to receive OLF filters for IPv4/IPv6 Prefix FECs from LSR1. This triggers LSR2 to construct an "OLF Capability" TLV in the same manner as described in section 4.4.1. The constructed "OLF Capability" is sent in a Capability message (with S-bit set to 1) towards LSR1. Upon receipt of this Capability message, LSR1 realizes that LSR2 is now capable to receive OLF filters for IPv4/IPv6 Prefix FEC types. As described in earlier section, LSR1 now proceeds by constructing "OLF Policy Status" using its configured filters for LSR2, and sends them in an LDP Notification message towards LSR2. Upon receipt of this message, LSR2 applies the received OLF policy and withdraws any label bindings corresponding to matching FEC (prefixes) that are no more permitted for advertisement. Later on, LSR1 can also update its OLF filters by pushing updates to LSR2 as/when any change in LSR1's OLF policy occurs.

Assuming both LSR1 and LSR2 are already engaged in OLF filtering in sender and receiver roles respectively for given FEC-Types. Now consider that LSR2 configuration is changed to remove "receive" capability for one FEC-Type (say IPv4 Prefix) from LSR1. This triggers LSR2 to construct an "OLF Capability" TLV that includes

only one OLF Capability Element corresponding to "IPv4 Prefix" FEC type. The constructed "OLF Capability" is sent in a Capability message (with S-bit set to 0) towards LSR1. Upon receipt of this Capability [withdrawal] message, LSR1 marks LSR2 as IPv4 Prefix FEC OLF "receive" incapable peer, and makes sure that no more OLF filter updates (via LDP Notification messages) are sent to LSR2. LSR2, after sending the Capability [withdrawal] message, now deletes any installed OLF filter corresponding to LSR1 for "IPv4 Prefix" FEC, and re advertises its entire label bindings database for "IPv4 Prefix" FEC to LSR1. Upon receipt of unwanted label bindings, LSR1 may chose to discard them by applying the filtering policy in inbound direction.

4.4.3. OLF Policy Updates

After successful negotiation of "OLF Capability" for a FEC-Type with the peer as the receiver and self as the sender, an LSR SHOULD now send its OLF policy to its peer via "OLF Policy Status" TLV in an LDP Notification message. The LSR MAY also update its OLF policy towards its peer by sending further updates, if/when its locally configuration/policy changes.

Consider LSR1 as sender and LSR2 as receiver of OLF filters for IPv4/IPv6 Prefix FEC types. After successful negotiation of OLF capabilities, LSR1 proceeds by sending its OLF filters towards LSR2 via LDP Notification message. LSR1 first constructs Status TLV and sets its status code to "OLF Status", and adds the "OLF Policy Status" TLV in the optional parameter section of the Notification message. The contents of "OLF Policy Status" TLV are constructed as set of OLF filters as defined by local configuration and policy for one or more OLF types. The sender MUST only include those OLF types in this TLV for which it has successfully negotiated the OLF capability with the peer. In our example, LSR1 constructs two OLF Elements for IPv4 and IPv6 Prefix FEC types. Each OLF Element is constructed with one ore more OLF Entries, as defined by or mapped to locally configured OLF policy corresponding to LSR2. LSR1 then sends the constructed "OLF Policy Status" TLV, alongwith Status TLV (with status set to "OLF Status") in a LDP Notification message to LSR2.

The receiver LDP speaker LSR2 MUST honor the receipt of this TLV in a Notification message because it had successfully negotiated the capability as the receiver for one or more OLF types. If an LDP speaker receives a "OLF Policy Status" TLV in a Notification message without prior OLF Capability(ies) exchange and negotiation, or if negotiated OLF Capability as sender-only role, it MUST ignore the received "OLF Policy Status" TLV, send a "Unknown TLV" Notification

back to the peer, and continue processing rest of the message. Similarly, LSR2 behaves the same way on receipt of this TLV in a Notification message with status code other than "OLF Status", and respond back with "Malformed TLV" Notification.

If the receiver LSR2 does not understand or does not support the FEC-Type (FEC Element type and/or Address Family) specified in an "OLF Element", it MUST respond with a LDP Notification with status code set to "Unknown FEC" or "Unsupported Address Family" as applicable, and abort processing of the entire message.

If LSR1's configured OLF policy changes, LSR1 sends further updates using "OLF Policy Status" in a LDP Notification message. Upon receipt of such an update for given FEC-Type, LSR2 treats this as an overwrite of the previously installed OLF filters corresponding to LSR1, and re-applies the policy. As the result of policy re-application, LSR2 advertises any new [matching] prefix being permitted now, and withdraws any previously advertised prefixes which are no longer permitted as per matching rules.

5. Address Prefix FEC OLF Type

Using the earlier OLF framework defined in this document, this section defines the OLF type for the "Address Prefix" FEC Element type. The OLF types for other FEC Element types are beyond the scope of this document.

The "Address Prefix FEC" OLF type allows one to express OLFs in terms of address prefixes. That is, it provides filtering based on address prefixes, including prefix length or range based matching.

To define an OLF for "Address Prefix FEC" type of given address family, the FEC-Elem-Type and Address-Family fields of an OLF Element are defined as follows:

FEC-Elem-Type: 0x2 ("Address Prefix")
Address-Family: 1 (IPv4) or 2 (IPv6)

Conceptually, an "Address Prefix FEC" OLF entry for a given Address Family consists of the fields <Action, Prefix Length, Prefix, Minlen, Maxlen>, and hence the "Address Prefix FEC" OLF entry within an "Address Prefix FEC" OLF element is encoded as follows:

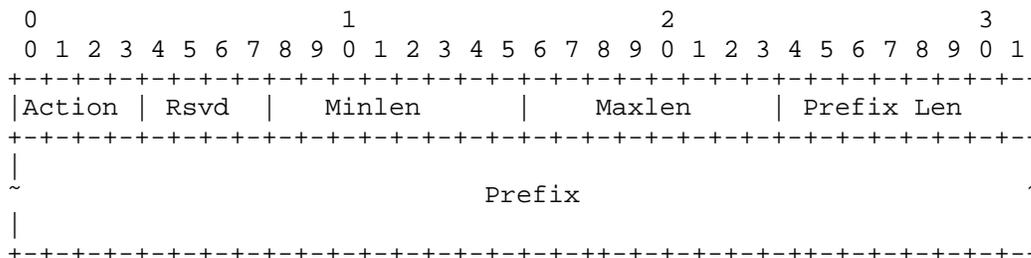


Figure 7: Address Prefix FEC OLF Entry

With reference to Fig 3, the first octet of the above OLF Entry belongs to the "Common part" and the rest of the fields belong to the "Type-specific part" (as defined for Address Prefix FEC Element type).

As per OLF Entry rules defined earlier, if the Action component of the entry specifies wildcard operation ("PERMIT-ALL"), then Address Prefix FEC OLF Entry does not specify any type-specific data (i.e. OLF entry size is 1 octet only).

The "Minlen" and "Maxlen" fields indicate respectively the minimum and the maximum prefix length in bits that is used for "matching". Either the Minlen or Maxlen field or both may have the value 0; this means that the value of the field is "unspecified". The Maxlen value must not be more than the maximum length (in bits) of a host address for the given address family.

The "Prefix Len" field indicates the length in bits of the address prefix. This field MUST NOT be specified as zero.

The "Prefix" field contains an address prefix encoded according to the given address family.

This document imposes that values of these fields MUST satisfy the following rule, assuming Minlen and Maxlen are specified:

$$0 < \text{Prefix Len} \leq \text{Minlen} \leq \text{Maxlen}$$

5.1. Matching Address Prefixes to OLF Entries

Consider an Address Prefix FEC OLF entry, and an IP route maintained by an LDP speaker in the form of <Prefix, Prefix Length>. Following are the matching rules defined for Address Prefix OLF specific matching.

- o The IP route is considered as "no match" to the OLF entry if the route prefix is neither more specific than, nor equal to, the <Prefix, Prefix Len> fields of the OLF entry.
- o When the IP route is either more specific than, or equal to, the <Prefix, Prefix Len> fields of the OLF entry, the route is considered as a match to the OLF entry only if the match conditions as listed in Table 2 are satisfied (where un-spec refers to a value of zero).

OLF Entry		Route Prefix
Minlen	Maxlen	Match Condition
un-spec.	un-spec.	Route.Prefix Len == OLF.Prefix Len
specified	un-spec.	Route.Prefix Len >= OLF.Minlen
un-spec.	specified	Route.Prefix Len <= OLF.Maxlen
specified	specified	Route.Prefix Len >= OLF.Minlen AND Route.Prefix Len <= OLF.Maxlen

Table 2: Address Prefix OLF Entry Matching Rules

- o When more than one Address Prefix OLF entry matches the route, the "first-match" rule applies. That is, the OLF entry that is specified (and processed) first in a given OLF update (among all the matching OLF entries) is considered as the sole match, and it would determine whether the route should be permitted or denied.

6. Operational Examples

6.1. Label Filtering at Area Border Router

A typical service provider core network is designed with two or more levels of IGP hierarchy. In OSPF parlance, a backbone area is connected to multiple islands of non-zero areas. Similarly, in an IS-IS network, core L2 areas are connected to L1 areas. When LDP is enabled in such a network, an ABR (or a L2 router) that connects multiple non-zero areas to the backbone will advertise LDP label bindings for all prefixes (non-zero area as well as backbone area). However, depending on the MPLS hierarchy, each ABR may want label bindings for only the backbone area prefixes. The OLF scheme specified in this document provides a mechanism to do so efficiently.

6.2. LSR with limited LIB size

Assume an LSR (LSR1) is not capable of storing all IPv4 label bindings from its peer (LSR2) in its IPv4 Label Information Base (LIB), and it is desirable to receive and store only handful of remote label bindings from its peer. One approach of solving this issue is to use Downstream on Demand mode of label distribution so that LSR2 does not send its entire label database unsolicitedly towards LSR1. Instead, LSR1 uses Label Request mechanics to request labels for [handful of] interested FECs from its peer LSR2. This approach has few drawbacks:

- a. This forces Downstream On Demand label distribution mode on both LSRs (LSR1 and LSR2) engaged in the session, although this mode is really required by LSR1 due to its limitation.
- b. The control plane signaling convergence for Downstream On Demand label distribution mode is slower than Downstream Unsolicited.

An alternate approach to meet LSR1 requirement is to use OLF mechanics while using Downstream Unsolicited distribution mode. In this approach, LSR1 and LSR2 will negotiate OLF Capability as sender/receiver respectively, and LSR1 will install OLF filters to limit the IPv4 label bindings sent by LSR2 to the only IPv4 prefixes in which LSR1 is interested in.

7. 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].

8. IANA Considerations

The document introduces following new protocol elements that require code point assignment by IANA:

- o "Outbound Label Filter Capability" TLV (requested code point: 0x50E)
- o "Outbound Label Filter Policy Status" TLV (requested code point: 0x50F)
- o "Outbound Label Filter Status" status code (requested code point: 0x00000050)

9. References

9.1. Normative References

- [RFC5036] Andersson, L., Menei, I., and Thomas, B., Editors, "LDP Specification", RFC 5036, September 2007.
- [RFC5561] Thomas, B., Raza, K., Aggarwal, S., Aggarwal, R., and Le Roux, JL., "LDP Capabilities", RFC 5561, July 2009.
- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC2119, March 1997.

9.2. Informative References

- [RFC5920] Fang, L. et al., "Security Framework for MPLS and GMPLS Networks", RFC 5920, July 2010.
- [RFC5291] Chen, E., Rekhter, Y., "Outbound Route Filtering Capability for BGP-4", RFC 5291, August 2008.
- [RFC5292] Chen, E., Sangli, S., "Address-Prefix-Based Outbound Route Filter for BGP-4", RFC 5292, August 2008.
- [RFC5918] Asati, R., Minei, I., and Thomas, B. "Label Distribution Protocol Typed Wildcard FEC", RFC 5918, August 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.
- [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.
- [P2MP-PW] Martini, L., Boutros, S., Sivabalan, S., Konstantynowicz, M., Del Vecchio, G., Nadeau, T., Jounay, F., Niger, P., Kamite, Y., Jin, L., Vigoureux, M., Ciavaglia, L., and Delord, S., "Signaling Root-Initiated Point-to-Multipoint Pseudowires using LDP", draft-ietf-pwe3-p2mp-pw-02.txt, Work in Progress, March 2011.

10. 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

Pradosh Mohapatra
Cisco Systems, Inc.
3750 Cisco Way,
San Jose, CA 95134, USA.
E-mail: pmohapat@cisco.com

MPLS Working Group
Internet Draft
Intended status: Informational
Expires: November 2011

R. Ram
D. Cohn
Orckit-Corrigent

M. Daikoku
KDDI

M. Yuxia
Y. Jian
ZTE Corp.

A. D'Alessandro
Telecom Italia

May 31, 2011

SD detection and protection triggering in MPLS-TP
draft-rkhd-mpls-tp-sd-03.txt

Abstract

This document describes guidelines for Signal Degrade (SD) fault condition detection at an arbitrary transport path (LSP or PW) and the usage of MPLS-TP fault management [3] for triggering protection switching as defined in the MPLS-TP survivability framework [2].

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/lid-abstracts.txt>

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

This Internet-Draft will expire in November 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.

Table of Contents

1. Introduction	3
2. Conventions used in this document	3
3. Signal Degradate and MPLS-TP protection switching	4
4. SD detection method	4
4.1. Guidelines for SD detection	4
4.2. Examples for SD detection methods	6
5. Transmission of link degradation fault indication	6
5.1. Lower layer Bit Error transmission	7
6. Handling of link degradation fault indication	7
7. Security Considerations	7
8. IANA Considerations	7
9. Acknowledgments	7
10. References	7
10.1. Normative References	7
10.2. Informative References	8

1. Introduction

Telecommunication carriers and network operators expect to replace aged TDM Services (e.g. legacy VPN services) provided by legacy TDM equipment by new VPN services provided by MPLS-TP equipment.

From a service level agreement (SLA) point of view, service quality and availability degradation are not acceptable, even after migration to MPLS-TP equipment.

In addition, from an operational point of view, comparable performance monitoring features to those provided by TDM networks are expected from MPLS-TP networks. For example, OAM maintenance points should be the same after TDM to MPLS-TP migration, as SLA revision is typically NOT feasible for telecommunication carriers and network operators.

MPLS-TP transport path (i.e. LSP,PW) resiliency actions such as protection switching can be triggered by fault conditions and external manual commands. Fault conditions include Signal Failure (SF) and Signal Degrade (SD). The SD condition could be detected at an intermediate link, based on lower layer indications or other sub-layer techniques.

Since the transport path protection switching is not necessarily managed by the transport entity that detects the SD condition, an indication of the link SD condition must be sent over the transport paths that traverse the affected link.

This document describes guidelines for SD detection by lower layers indication, and a mechanism for relaying the degraded transport path condition to the network element handling the protection switching at the appropriate transport path level.

2. Conventions used in this document

BER: Bit Error Rate

LSP: Label Switched Path

LSR: Label Switching Router

MEP: Maintenance End Point

MPLS: Multi-Protocol Label Switching

MPLS-TP: MPLS Transport Profile

OAM: Operations, Administration and Maintenance

OTN: Optical Transport Network

PCS: Physical Coding Sublayer

SF: Signal Failure

SD: Signal Degrade

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

3. Signal Degrade and MPLS-TP protection switching

Network survivability, as defined in [2], is the ability of a network to recover traffic delivery following failure or degradation of network resources. [5] defines an LSP protection mechanism and state machine that handles SF, SD and operator manual commands.

4. SD detection method

4.1. Guidelines for SD detection

Signal degrade is a transport path condition in which the expected quality of transport service delivery is not provided. The signal degrade condition can be used by operators to detect different types of failures, especially those with slow externalization such as optical device aging (e.g. photo detector and laser diode in line amplifier, transponder or SFP), transmission medium external impairment (e.g. temperature or pressure fluctuation, fiber elongation), and time-variable optical impairments in fiber (e.g. chromatic dispersion, polarization mode dispersion).

Signal degrade condition in a transport path is derived from bit error detection in the traversed links.

Bit errors in a link are caused by the following phenomena:

1. Physical conditions such as bad electrical connections, low received optical power, dispersion effects.

2. Non-physical conditions such as network congestion, CPU overload, selective packet discard, packet processing error.

The common basis for the guidelines set forth in this section is that the SD condition SHOULD reflect only physical error conditions in the traversed links, without any influence from non-physical conditions.

The following conditions SHOULD be met by the signal degrade condition detection mechanism:

- o Method for determining signal degrade MUST NOT affect the services transmitted over the transport path (e.g. add delay or jitter to real-time traffic)
- o Criterion for determining signal degrade MUST be agnostic to the length of transmitted frames
- o Criterion for determining signal degrade MUST be agnostic to the transmission rate of transmitted frames
- o Criterion for determining signal degrade MUST be agnostic to the type of service carried by the transmitted frames
- o Criterion for determining signal degrade MUST be agnostic to the traffic class of transmitted frames
- o Criterion for determining signal degrade MUST be agnostic to drop-precedence marking of transmitted frames
- o Criterion for determining signal degrade MUST be agnostic to congestion
- o Criterion for determining signal degrade SHOULD be able to detect low error levels (e.g. BER of $10E-8$)
- o Criterion for determining signal degrade SHOULD have low misdetection probability
- o Criterion for determining signal degrade SHOULD have low false alarm probability
- o Criterion for determining signal degrade SHOULD be agnostic to number of transport paths (LSPs and PWs) transported over the transmission link
- o Signal degrade conditions MUST be monitored by the lowest server layer or sub-layer that is not terminated between monitoring points

- o Method for determining signal degrade SHOULD NOT require transmission of additional packets
- o Method for determining signal degrade SHOULD allow to localize links that contribute to signal degrade
- o Method for determining signal degrade MUST be able to exit signal degrade condition when error rate returns to normal condition
- o Method for determining signal degrade condition MUST be scalable

4.2. Examples for SD detection methods

- o A Server MEP [4] related to SONET or SDH sub-layers can determine SD condition based on error indication from parity information in the path overhead.
- o A Server MEP related to OTN sub-layer can determine SD condition based on error indications from Forward-Error-Correction functionality inherent in encapsulation.
- o A Server MEP related to 10GE PCS sub-layer can determine SD condition based on rate of errored 66-bit block headers. (a.k.a. symbol errors)
- o A Server MEP related to 1GE PCS sub-layer can determine SD condition based on rate of 10-bit code violations dispersion errors.

As specified in section 4.1, these examples assume that the layer carrying the information used for SD detection is not terminated by non-MPLS-TP-LSR entities (e.g. media converter).

5. Transmission of link degradation fault indication

When SD condition is detected, a link degradation fault indication [3] SHOULD be transmitted over affected transport paths, in the downstream direction from the detection point. The link degradation indication will be transmitted immediately following the detection and periodically until the SD condition is removed. The messages will be terminated and handled by the downstream client MEP.

The encapsulation and mechanism defined in [3] is suitable for transmission of link degradation fault indication. It is RECOMMENDED that [3] will include this definition in future work.

5.1. Lower layer Bit Error transmission

There are scenarios where the lower layer bit error rate in each of the links traversed by the transport path is below the SD threshold, while the accumulated end-to-end BER on the LSP is above the threshold. This is possible in lower layer technologies where errored information is dropped, so errors in one link will not be detected by LSRs downstream of this link. An example of such a situation is when an LSP is carried over multiple Ethernet links, and each link drops errored Ethernet frames.

To enable SD detection in such scenarios, LSRs MAY optionally include the measured BER in the link degradation fault indication message. The client MEP may then receive multiple link degradation fault indication messages from different LSRs. When this occurs, the client MEP SHOULD compare the sum of the received BER values with the SD threshold to decide on the LSP SD condition.

6. Handling of link degradation fault indication

LSR behavior upon receiving link degradation fault indication is out of the scope of this document.

SD condition processing and prioritization for protection triggering is out of the scope of this document.

SD clear condition processing and prioritization for protection triggering is out of the scope of this document.

7. Security Considerations

To be added in a future version of the document.

8. IANA Considerations

<N/A>

9. Acknowledgments

The editors gratefully acknowledge the contributions of Amir Halperin and Shachar Katz.

10. References

10.1. Normative References

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

10.2. Informative References

- [2] Sprecher,N., and Farrel,A., "Multiprotocol Label Switching Transport Profile Survivability Framework", draft-ietf-mpls-tp-survive-fwk-06(work in progress), June 2010
- [3] Swallow,G., Fulignoli,A., Vigoureux,M., Boutros,S., and Ward,D., "MPLS Fault Management OAM", draft-ietf-mpls-tp-fault-04 (work in progress), April 2011
- [4] Busi,I. and Allan,D., "MPLS-TP OAM Framework", draft-ietf-mpls-tp-oam-framework-11 (work in progress), February 2011
- [5] Bryant,S., Osborne,E., Weingarten,Y., Sprecher,N., Fulignoli,A., "MPLS-TP Linear Protection", draft-ietf-mpls-tp-linear-protection-06 (work in progress), March 2011

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

Authors' Addresses

Rafi Ram
Orckit-Corrigent
126 Yigal Alon St.
Tel Aviv
Israel

Email: rafir@orckit.com

Daniel Cohn
Orckit-Corrigent
126 Yigal Alon St.
Tel Aviv
Israel

Email: danielc@orckit.com

Masahiro Daikoku
KDDI
3-10-10, Iidabashi, Chiyoda-ku,
Tokyo
Japan

Email: ms-daikoku@kddi.com

Ma Yuxia
ZTE Corp.
China

Email: ma.yuxia@zte.com.cn

Yang Jian
ZTE Corp.
China

Email: yang.jian90@zte.com.cn

Alessandro D'Alessandro
Telecom Italia
Italy

Email: alessandro.dalessandro@telecomitalia.it

Contributors

Amir Halperin

Shachar Katz

MPLS
Internet-Draft
Intended status: Standards Track
Expires: January 5, 2012

C. Villamizar, Ed.
Infinera Corporation
July 4, 2011

Multipath Extensions for MPLS Traffic Engineering
draft-villamizar-mpls-tp-multipath-te-extn-00

Abstract

Extensions to OSPF-TE, ISIS-TE, and RSVP-TE are defined in support of carrying LSP with strict packet ordering requirements over multipath and carrying LSP with strict packet ordering requirements within LSP without violating requirements to maintain packet ordering. LSP with strict packet ordering requirements include MPLS-TP LSP.

OSPF-TE and ISIS-TE extensions defined here indicate node and link capability regarding support for ordered aggregates of traffic, multipath traffic distribution, and abilities to support multipath load distribution differently per LSP.

RSVP-TE extensions either identifies an LSP as requiring strict packet order, or identifies an LSP as carrying one or more LSP that requires strict packet order at a given depth in the label stack, or identifies an LSP as having no restrictions on packet ordering except the restriction to avoid reordering microflows. In addition an extension indicates whether the first nibble of payload will reliably indicate whether payload is IPv4, IPv6, or other type of payload, most notably pseudowire using a pseudowire control word.

Status of this Memo

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

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

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

This Internet-Draft will expire on January 5, 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	4
1.1.	Architecture Summary	4
1.2.	Requirements Language	5
1.3.	Definitions	5
2.	Protocol Extensions	5
2.1.	Multipath Node Capability sub-TLV	6
2.2.	Multipath Link Capability sub-TLV	9
2.3.	Contained Ordered Aggregate Attributes TLV	9
2.4.	LSP Multipath Attributes TLV	11
3.	Multipath Extension Protocol Mechanisms	12
3.1.	OSPF-TE and ISIS-TE Advertisement	12
3.1.1.	Node Capability Advertisement	12
3.1.2.	Link Capability Advertisement	13
3.1.3.	Setting Max Depth and IP Depth	13
3.1.4.	Hierarchical LSP Link Advertisement	13
3.2.	RSVP-TE LSP Attributes	14
3.2.1.	LSP Attributes for Ordered Aggregates	14
3.2.2.	LSP Attributes for Ordered Aggregates	15
3.2.3.	Attributes for LSP without Packet Ordering	15
3.3.	Path Computation Constraints	16
3.3.1.	Link Multipath Capabilities and Path Computation	16
3.3.1.1.	Path Computation with Ordering Constraints	17
3.3.1.2.	Path Computation with No Ordering Constraint	17
3.3.1.3.	Path Computation for MPLS containing MPLS-TP	17
3.3.2.	Link IP Capabilities and Path Computation	18
3.3.2.1.	LSP without Packet Ordering Requirements	18
3.3.2.2.	LSP with Ordering Requirements	19
3.3.2.3.	LSP containing LSP with Ordering Requirements	19
3.3.3.	Link Depth Limitations and Path Computation	19
4.	Backwards Compatibility	20
4.1.	Legacy Multipath Behavior	20
4.2.	Networks without Multipath Extensions	21
4.2.1.	Networks with MP Capability on all Multipath	21
4.2.2.	Networks with OA Capability on all Multipath	22
4.2.3.	Legacy Networks with Mixed MP and OA Links	22
4.3.	Transition to Multipath Extension Support	22
4.3.1.	Simple Transitions	23
4.3.2.	More Challenging Transitions	23
5.	IANA Considerations	23
6.	Security Considerations	24
7.	References	24
7.1.	Normative References	24
7.2.	Informative References	24
	Author's Address	26

1. Introduction

Today the requirement to handle large aggregations of traffic, can be handled by a number of techniques which we will collectively call multipath. Multipath is very similar to composite link as defined in [ITU-T.G.800], except multipath specifically excludes inverse multiplexing. Some types of LSP, including but potentially not limited to MPLS-TP LSP, require strict packet ordering.

Requirements to carry MPLS-TP LSP over multipath and a framework giving a number of methods to do so are defined in [I-D.villamizar-mpls-tp-multipath]. This document defines protocol extensions which provide a means of supporting MPLS as a server layer for MPLS-TP, or to carry MPLS-TP directly over a network which makes use of multipath.

1.1. Architecture Summary

Advertisements in a link state routing protocol, such as OSPF or ISIS, support a topology map known as a link state database (LSDB). When traffic engineering information is included in the LSDB the topology map is known as a TE-LSDB or traffic engineering database (TED).

A common MPLS LSP path computation is known as a constrained shortest path first computation (CSPF) (see [RFC3945]). Other algorithms may be used for path computation. Constraint-based routing was first introduced in [RFC2702]).

OSPF-TE or ISIS-TE extensions are defined in Section 2.1 and Section 2.2. OSPF-TE or ISIS-TE advertisements serve to populate the TE-LSDB and provide the basis for constraint-based routing path computation. Section 3.1 describes the use of OSPF-TE or ISIS-TE multipath extensions in routing advertisements.

RSVP-TE extensions are defined in Section 2.3 and Section 2.4. Section 3.2 describes the use of RSVP-TE extensions in setting up LSP including signaling constraints on LSP which contain other LSP which specify RSVP-TE extensions.

Section 3.3 describes the constraints on LSP path computation imposed by the advertised ordered aggregate and multipath capabilities of links. Section 3.3.2 describes the constraints on LSP path computation imposed by link advertisements regarding use of IP headers in multipath traffic distribution. Section 3.3.3 describes the impact of label stack depth limitations.

1.2. Requirements Language

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

1.3. Definitions

Please refer to [I-D.villamizar-mpls-tp-multipath].

Ordered Aggregate (OA)

An ordered aggregate (OA) requires that packets be delivered in the order in which they were received. Please refer to [RFC3260].

Microflow

A microflow is a single instance of an application-to-application flow. Please refer to [RFC2475]. Reordering packets within a microflow can cause service disruption. Please refer to [RFC2991].

Multipath Traffic Distribution

Multipath traffic distribution refers to the mechanism which distributes traffic among a set of component links or component lower layer paths which together comprise a multipath. No assumptions are made about the algorithms used in multipath traffic distribution. This document only discusses constraints of the type of information which can be used as the basis for multipath traffic distribution in specific circumstances.

The phrase "strict packet ordering requirements" refers to the requirement to deliver all packet in the order that they were received. The absence of strict packet ordering requirements does not imply total absence of packet ordering requirements. The requirement to avoid reordering traffic within any given microflow, as described in [RFC2991] applies to all traffic aggregates including all MPLS LSP.

2. Protocol Extensions

This section defined protocol extensions to OSPF-TE, ISIS-TE, and RSVP-TE to address requirements described in [I-D.villamizar-mpls-tp-multipath].

Two capability sub-TLV are added to two TLV that are used in both OSPF-TE and ISIS-TE. The Multipath Node Capability sub-TLV is added to the Node Attribute TLV (Section 2.1) The Multipath Link Capability sub-TLV is added to the Link Identification TLV (Section 2.2).

Two TLV are added to the LSP_ATTRIBUTES object defined in [RFC5420].

2.1. Multipath Node Capability sub-TLV

The Node Attribute TLV is defined in [RFC5786]. A new sub-TLV, the Multipath Node Capability sub-TLV, is defined for inclusion in the Node Attribute TLV.

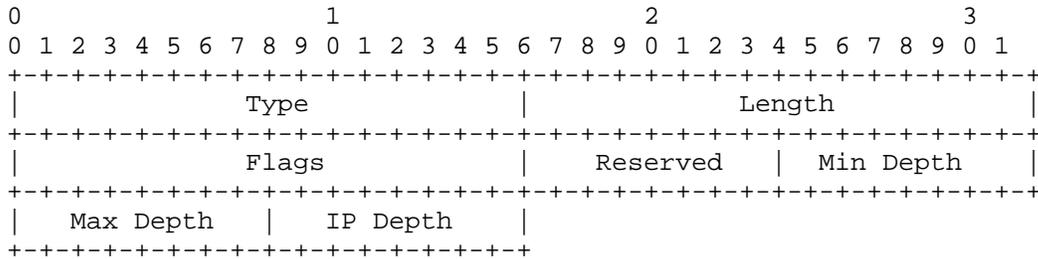


Figure 1: Multipath Capability Sub-TLV

The fields in the Multipath Capability sub-TLV are defined as follows.

Type

The Type field is assigned a value of IANA-TBD-1. The Type field is a two octet value.

Length

The Length field indicates the length of the sub-TLV in octets, excluding the Type and Length fields. The Length field is a two octet value.

Flags

The Flags field is a two octet (16 bit) value. The following single bit fields are assigned within this value, starting at the most significant bit, which is the bit transmitted first.

0x8000 Ordered Aggregate Enabled

Setting the Ordered Aggregate Enabled bit indicates that an LSP can be carried as an Ordered Aggregate Enabled on one or more links.

0x4000 Multipath Enabled

Setting the Multipath Enabled bit indicates that an LSP can be spread across component links at one or more multipath links.

0x2000 IPv4 Enabled Multipath

Setting the IPv4 Enabled Multipath bit indicates that the IPv4 header information can be used in multipath load balance. The Multipath Enabled bit must be set if the IPv4 Enabled Multipath bit is set.

0x1000 IPv6 Enabled Multipath

Setting the IP bit indicates that the IPv6 header information can be used in multipath load balance. The Multipath Enabled bit must be set if the IPv6 Enabled Multipath bit is set.

0x0800 UDPIIPv4 Multipath

Setting the UDPIIPv4 Multipath bit indicates that the UDP port numbers carried in UDP over IPv4 can be used in multipath load balance. The IPv4 Enabled Multipath bit must be set if UDPIIPv4 Multipath is set. If the IPv4 Enabled Multipath bit is set and the UDPIIPv4 Multipath bit is clear, then only source and destination IP addresses are used.

0x0400 UDP/IPv6 Multipath

Setting the UDP/IPv6 Multipath bit indicates that the UDP port numbers carried in UDP over IPv6 can be used in multipath load balance. The IPv6 Enabled Multipath bit must be set if UDP/IPv6 Multipath is set. The IPv4 Enabled Multipath bit must be set if UDPIIPv4 Multipath is set. If the IPv6 Enabled Multipath bit is set and the UDP/IPv6 Multipath bit is clear, then only source and destination IP addresses are used.

0x0200 TDPIIPv4 Multipath

Setting the TDPIIPv4 Multipath bit indicates that the TCP port numbers carried in TCP over IPv4 can be used in multipath load balance. The IPv4 Enabled Multipath bit must be set if TDPIIPv4 Multipath is set. If the IPv4 Enabled Multipath bit is set and the TDPIIPv4 Multipath bit is clear, then only source and destination IP addresses are used.

0x0100 TCP/IPv6 Multipath

Setting the TCP/IPv6 Multipath bit indicates that the TCP port numbers carried in TCP over IPv6 can be used in multipath load balance. The IPv6 Enabled Multipath bit must be set if TCP/IPv6 Multipath is set. The IPv4 Enabled Multipath bit must be set if TDPIIPv4 Multipath is set. If the IPv6 Enabled Multipath bit is set and the TCP/IPv6 Multipath bit is clear, then only source and destination IP addresses are used.

0x0080 Default to Multipath

Setting the Default to Multipath bit indicates that for an LSP which does not signal a desired behavior the traffic for that LSP will be spread across component links at one or more multipath links. If the Default to Multipath bit is not set, then an LSP which does not signal otherwise will be treated as an ordered aggregate.

0x0040 Default to IP/MPLS Multipath

Setting the Default to IP/MPLS Multipath indicates that for an LSP which does not signal a desired behavior, the IP header information will be used in the multipath load distribution. If the Default to IP/MPLS Multipath is clear it indicates that the the IP header information will not be used by default.

0x0020 Variable Depth Multipath

Setting the Variable Depth Multipath bit indicates that when multipath is enabled for a given LSP, the stack depth beyond which multipath will not extract information for use in the multipath load distribution can be set on a per LSP basis.

0x0010 IP Optioal Multipath

Setting the IP Optioal Multipath bit indicates that when multipath is enabled for a given LSP, whether the IP header information is used in the multipath load distribution can be set on a per LSP basis.

The remaining bits in the Flags field are reserved.

Reserved

The Reserved field MUST be set to zero and MUST be ignored unless implementing an extension.

Min Depth

The Min Depth field indicates the minimum stack depth which can be supported in the multipath traffic distribution. For links which are not PSC LSP, the Min Depth field is set to zero. For FA advertised for PSC LSP, this field will be set to one or more. See Section 3.1.4 for further details.

Max Depth

The Max Depth field is a one octet field indicating the maximum label stack depth beyond which the multipath load distribution cannot make use of further label stack entries.

IP Depth

The IP Depth field is a one octet field indicating the maximum label stack depth beyond which the multipath load distribution cannot make use of IP information.

The reserved bits in the Flags field and the Reserved field MUST be set to zero and MUST be ignored unless implementing an extension which redefines one or more of the reserved bits. Any further extension which redefines one or more reserved Flags bit should maintain backwards compatibility with prior implementations.

2.2. Multipath Link Capability sub-TLV

The Link Identification TLV is defined in [RFC3471]. The Link Identification TLV is updated in [RFC4201] to support a form of multipath known as Link Bundling.

A new sub-TLV, the Multipath Link Capability sub-TLV, is defined here. The Multipath Link Capability sub-TLV is optionally included in the Link Identification TLV. The format of the Multipath Link Capability sub-TLV is identical to the Multipath Node Capability sub-TLV defined in Section 2.1, with one exception. In the Multipath Link Capability sub-TLV the Type field value is IANA-TBD-2.

If a Multipath Link Capability sub-TLV is advertised for any link, then a Multipath Node Capability sub-TLV MUST be advertised for the node.

2.3. Contained Ordered Aggregate Attributes TLV

The LSP_ATTRIBUTES object is defined in [RFC5420]. A new Contained Ordered Aggregate Attributes TLV is defined for the LSP_ATTRIBUTES object. The TLV Type is IANA_TBD_3. The format is described below.

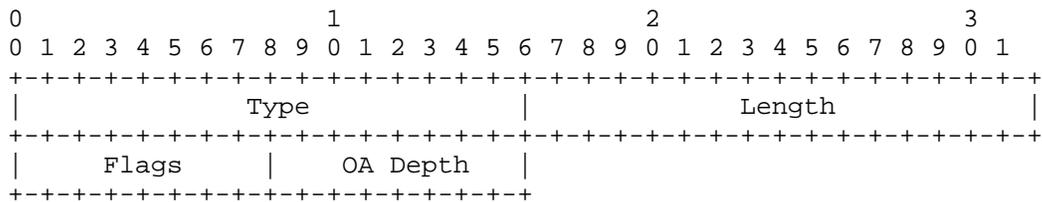


Figure 2: Contained Ordered Aggregate Attributes TLV

The fields in the Contained Ordered Aggregate Attributes TLV are defined as follows.

Type

The Type field is assigned a value of IANA-TBD-3. The Type field is a two octet value.

Length

The Length field indicates the length of the sub-TLV in octets, excluding the Type and Length fields. The Length field is a two octet value.

Flags

The Flags field is a three octet (24 bit) value. The following single bit fields are assigned within this value, starting at the most significant bit, which is the bit transmitted first.

0x80 IP Multipath Allowed

Setting the IP Multipath Allowed bit indicates that it is safe to enable the use of a potential IP payload in the multipath traffic distribution.

0x40 May Contain IPv4

Setting the May Contain IPv4 bit indicates that IPv4 traffic may be contained within this LSP.

0x40 May Contain IPv6

Setting the May Contain IPv6 bit indicates that IPv6 traffic may be contained within this LSP.

The remaining bits in the Flags field are reserved.

OA Depth

The OA Depth field is set as follows

0 An OA Depth value of zero indicates that no ordered aggregates are carried within the LSP.

1 An OA Depth value of one indicates that the LSP is an ordered aggregate of traffic (the LSP requires strict ordering of packets).

>1 An OA Depth value greater than one indicates that the LSP does not have strict packet ordering requirements but contains ordered aggregates at the label stack depth indicated.

The reserved bits in the Flags field MUST be set to zero and MUST be ignored unless implementing an extension which redefines one or more of the reserved bits. Any further extension which redefines one or more reserved Flags bit should maintain backwards compatibility with prior implementations.

2.4. LSP Multipath Attributes TLV

The LSP_ATTRIBUTES object is defined in [RFC5420]. A new LSP Multipath Attributes TLV is defined for the LSP_ATTRIBUTES object. The TLV Type is IANA_TBD_4. The format is described below.

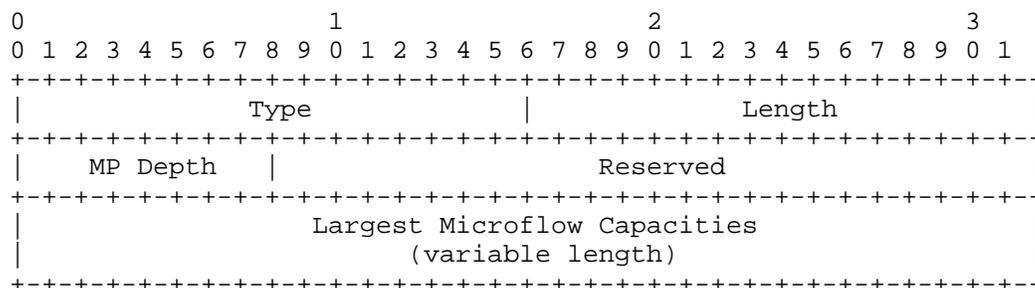


Figure 3: LSP Multipath Attributes TLV

The fields in the LSP Multipath Attributes TLV are defined as follows.

Type

The Type field is assigned a value of IANA-TBD-4. The Type field is a two octet value.

Length

The Length field indicates the length of the sub-TLV in octets, excluding the Type and Length fields. The Length field is a two octet value.

MP Depth

The MP Depth field indicates the depth at which the Largest Microflow Capacities parameters are applicable.

Largest Microflow Capacities

The Largest Microflow Capacities field contains, one, two, or three IEEE 32 bit floating point values. Each value is a capacity expressed in bytes per second.

Largest LSE Microflow

The first value, the Largest LSE Microflow, is the capacity of the largest microflow if only the label stack entries are used in multipath traffic distribution. If a Largest LSE Microflow is not included, the LSP bandwidth request MUST be used.

Largest IP Microflow

The second value, the Largest IP Microflow, if present, is the capacity of the largest microflow if the label stack entries and any potential IP source and destination address are used in multipath traffic distribution. If the Largest IP Microflow is not included, the value of the Largest LSE Microflow MUST be used.

Largest L4 Microflow

The third, the Largest L4 Microflow, if present, is the capacity of the largest microflow if the label stack entries and any potential IP addresses and TCP or UDP port numbers are used in multipath traffic distribution. If a Largest L4 Microflow is not included, the value of the Largest IP Microflow MUST be used.

The Reserved field MUST be set to zero and MUST be ignored unless implementing an extension which redefines all or part of this field. Any further extension which redefines all or part of this field should maintain backwards compatibility with prior implementations.

If one or more LSP Multipath Attributes TLV is present, a Contained Ordered Aggregate Attributes TLV MUST be present. There SHOULD be no more than one LSP Multipath Attributes TLV for any value of the MP Depth field in any given LSP_ATTRIBUTES object. If additional LSP Multipath Attributes TLV are encountered they MUST be ignored.

The value of the MP Depth field must be greater than zero and less than the value of the OA Depth field in the Contained Ordered Aggregate Attributes TLV, unless the OA Depth field is set to zero.

3. Multipath Extension Protocol Mechanisms

3.1. OSPF-TE and ISIS-TE Advertisement

Every node MUST advertise exactly one Multipath Node Capability sub-TLV and may advertise zero or more Multipath Link Capability sub-TLV as needed.

3.1.1. Node Capability Advertisement

Every LSR which is adjacent to one or more multipath link MUST advertise a Multipath Node Capability sub-TLV (see Section 2.1). The capabilities advertised for the node SHOULD reflect the capabilities of the majority of multipath links adjacent to the node.

3.1.2. Link Capability Advertisement

For all of the links whose capability does not exactly match the Multipath Node Capability sub-TLV advertised by that same LSR, the LSR MUST advertise a Multipath Link Capability sub-TLV (see Section 2.2).

For all of the links whose capability does exactly match the Multipath Node Capability sub-TLV advertised by that same LSR, the LSR SHOULD NOT advertise a Multipath Link Capability sub-TLV (see Section 2.2). In this case the Multipath Link Capability sub-TLV is redundant, but harmless.

3.1.3. Setting Max Depth and IP Depth

The Max Depth and IP Depth field are intended to capture architectural limits. Most forwarding hardware will only use a limited number of labels in the multipath traffic distribution. This limit is reflected in the Max Depth field. Most forwarding hardware will limit the number of labels that it will look past before looking for an IP header to be used in the multipath traffic distribution. This limit is reflected in the IP Depth field.

3.1.4. Hierarchical LSP Link Advertisement

A PSC LSP, as defined in [RFC4206] and updated in [RFC6107], may carry other LSP. When signaling a PSC LSP that is expected to carry MPLS-TP LSP or other LSP with strict packet reordering requirements at some label depth, the minimum label stack depth at which an LSP with strict packet reordering requirements can be carried must be signaled.

A tradeoff must be considered. If allowed to look deep into the label stack, multipath traffic distribution is able to distribute traffic more evenly. It is impractical to carry LSP very deep in the stack solely for this purpose. When carrying LSP that has strict packet order requirements, the depth at which these LSP will be carried is a network design tradeoff.

For example, if an MPLS-TP LSP is signaled as a PSC LSP, the the depth needs to be limited to one. To directly carry an LSP that requires strict packet ordering, the depth needs to be limited to two, where the two label stack entries are for the MPLS LSP with no packet order requirements other than for microflows and the contained LSP with strict packet order requirements.

When the Forwarding Adjacency (FA) is advertised, the depth at which the label stack will inspected indicated in signaling at setup time

is expressed in the Min Depth field of the Multipath Link Capability sub-TLV.

The advertised Max Depth and IP Depth of a PSC LSP must be one less than the minimum of the Max Depth and IP Depth of any link that the PSC LSP traverses. The Max Depth and IP Depth are considered independently of each other.

3.2. RSVP-TE LSP Attributes

All LSP SHOULD advertise a Contained Ordered Aggregate Attributes TLV. LSP with strict packet order requirements MUST set the OA Depth field to one to indicate that the LSP MUST be treated as ordered aggregate. LSP which do not have strict packet order requirements MUST only carry LSP whose requirements are reflected in the containing LSP Contained Ordered Aggregate Attributes TLV and LSP Multipath Attributes TLVs or the containing LSP MUST either resignal appropriate TLVs using make-before-break semantics, or the contained LSP signaling must be rejected with a PathErr message. Three cases are described in the following subsections.

3.2.1. LSP Attributes for Ordered Aggregates

The Flags field in the Contained Ordered Aggregate Attributes TLV MUST be set as follows.

1. If the LSP may directly contain IPv4 traffic, then the May Contain IPv4 bit in the Flags field MUST be set.
2. If the LSP may directly contain IPv6 traffic, then the May Contain IPv6 bit in the Flags field MUST be set.
3. If the LSP contains an LSP which has the May Contain IPv4 bit in the Flags field then the May Contain IPv4 bit in the Flags field MUST be set.
4. If the LSP contains an LSP which has the May Contain IPv6 bit in the Flags field then the May Contain IPv6 bit in the Flags field MUST be set.
5. If the LSP may contain psuedowires that do not use a pseudowire control word, or may contain IPv4 or IPv6 traffic, then the IP Multipath Allowed bit in the Flags field MUST be cleared.
6. If the LSP is known not to contain psuedowires that do not use a pseudowire control word, and is known not to contain IPv4 or IPv6 traffic, then the IP Multipath Allowed bit in the Flags field SHOULD be set unless disallowed due to a contained LSP.

7. If the LSP is known not to contain pseudowires that do not use a pseudowire control word, and does not have strict packet ordering requirements, then the IP Multipath Allowed bit in the Flags field SHOULD be set unless disallowed due to a contained LSP.
8. If the IP Multipath Allowed bit in the Flags is set and the LSP has strict packet order requirements, the May Contain IPv4 and May Contain IPv6 MUST be clear.
9. If the LSP contains any LSP with the IP Multipath Allowed bit in the Flags field clear, then the IP Multipath Allowed bit in the Flags field MUST be clear.
10. If the LSP contains any LSP with either the May Contain IPv4 bit or the May Contain IPv6 bit in the Flags field set, and the containing LSP has strict packet ordering requirements, then the IP Multipath Allowed bit in the Flags field MUST be clear.

If the LSP does not contain other LSP, then it can only contain pseudowire that terminate on that LSR. If the LSP does not contain other LSP, then it should be known whether the LSP is used in an IP LER capacity. Therefore, when a LSP does not contain other LSP, it should always be possible to accurately set the Flags field in the Contained Ordered Aggregate Attributes TLV.

See Section 4 for guidelines for handling LSP which contain LSP that do not have a Contained Ordered Aggregate Attributes TLV. The most conservative approach in this case is to clear the IP Multipath Allowed bit and set the May Contain IPv4 bit and the May Contain IPv6 bit, however this may not always be necessary.

3.2.2. LSP Attributes for Ordered Aggregates

An LSP with strict packet order requirements SHOULD always include a Contained Ordered Aggregate Attributes TLV. The OA Depth field MUST be set to one. The LSP SHOULD NOT include any LSP Multipath Attributes TLV.

3.2.3. Attributes for LSP without Packet Ordering

If an LSP does not have strict packet order constraints, then the LSR_ATTRIBUTE object SHOULD always include a Contained Ordered Aggregate Attributes TLV. The OA Depth MUST be either set to zero or set to a configured value that is greater than one.

If the OA Depth is set to a configured value, then any setup attempt for a contained LSP with a depth greater than or equal to that value

SHOULD be rejected and a PathErr message sent. Otherwise, if a setup attempt for a contained LSP with a depth greater than the current value included in the containing LSP OA Depth field, then the containing LSP MUST be rerouted with a OA Depth field value greater than any of the contained OA Depth field values.

The values in the LSP Multipath Attributes TLV may be constrained to upper limits by configuration. If an attempt to setup a contained LSP would result in exceeding one of these limits, then the LSR SHOULD reject the signaling attempt and send a PathErr message.

If an LSP does not have strict packet order constraints, then the LSR_ATTRIBUTE object MAY contain one or more LSP Multipath Attributes TLV. If the LSP does not contain any other LSP, then one LSP Multipath Attributes TLV MAY be contained, with the MP Depth field set to one. In this case, the Largest LSE Microflow in the Largest Microflow Capacities field MUST be set to the requested bandwidth of the LSP. The optional Largest IP Microflow and Largest L4 Microflow MAY be included and set to configured values.

If an LSP that does not have strict packet order constraints contains other LSP, then the LSP Multipath Attributes TLV advertised by the set of contained LSP MUST be used to set the LSP Multipath Attributes TLV Largest Microflow Capacities values for LSP Multipath Attributes TLV. The value of Largest LSE Microflow, Largest IP Microflow, and Largest L4 Microflow in the LSP Multipath Attributes TLV of the containing LSP with an MP Depth of N cannot be less than the maximum effective value of the same parameter for any contained LSP Multipath Attributes TLV with an MP Depth value of N-1.

3.3. Path Computation Constraints

The RSVP-TE extensions provides a set of requirements to be met by the links which the LSP is to traverse. This set of requirements also serves as the basis for path computation constraints and for admission control constraints.

3.3.1. Link Multipath Capabilities and Path Computation

Three cases are considered. An LSP may have strict ordering constraints. An MPLS-TP LSP is an example of an LSP with strict ordering constraints. This first type of LSP is covered in Section 3.3.1.1. An LSP may have no ordering constraints at all other than the constraint that microflows cannot be reordered. This second case is covered in Section 3.3.1.2. The remaining case is where an LSP has no ordering constraints but carries traffic for other LSP which do have ordering constraints. This third case is covered in Section 3.3.1.3.

3.3.1.1. Path Computation with Ordering Constraints

For an MPLS-TP LSP or other LSP with a strict packet ordering constraint, any link for which the Ordered Aggregate Enabled bit is not set must be excluded from the path computation. If the Default to Multipath bit is set on a link, then extensions to RSVP-TE to indicate a requirement to maintain packet order must be used in signaling to override the default.

3.3.1.2. Path Computation with No Ordering Constraint

For an MPLS LSP which has no constraint on packet ordering except that microflows must remain in order and does not contain other LSP with ordering constraints, any link for which the Multipath Enabled bit is set SHOULD use an available bandwidth taken from the "Unreserved Bandwidth" rather than the "Maximum LSP Bandwidth" (see [RFC4201]).

For most LSP, the bandwidth requirement of the largest microflow is not known but an upper bound is known. For example if the LSP aggregates PW or other LSP of no more than some maximum capacity or LSP which have signaled a microflow upper bound, then an upper bound on the largest microflow is known. If this upper bound exceeds the "Maximum LSP Bandwidth" of a given link, then that link SHOULD be excluded from the path computation.

3.3.1.3. Path Computation for MPLS containing MPLS-TP

To carry LSP which have strict packet ordering requirements within LSP that do not have strict packet ordering requirements, the label stack depth at which multipath traffic distribution is allowed to take information must be limited. To set up such an LSP, the minimum label stack depth at which an MPLS-TP LSP or other LSP with strict ordering constraints will be carried must be known.

For links which have the Variable Depth Multipath bit clear, the MPLS LSP MUST be treated as if the containing LSP has ordering constraints, unless the Max Depth for the link is equal to the minimum label stack depth at which an MPLS-TP LSP or other LSP with strict ordering constraints will be carried. If the LSP is treated as if the containing LSP has ordering constraints, bandwidth constraints MUST be applied as described in Section 3.3.1.1. Failing to do so would violate the ordering constraints of contained LSP.

For links which have the Variable Depth Multipath bit set, constraints may be applied to links in the path computation as described in Section 3.3.1.2. The minimum label stack depth at which an MPLS-TP LSP or other LSP with strict ordering constraints is carried MUST be

signaled when the LSP is set up.

The minimum label stack depth at which an MPLS-TP LSP or other LSP with strict ordering constraints is carried limits the multipath load balance and therefore requires an additional constraint. For LSP that cannot be further subdivided using information in IP headers below the MPLS stack, those LSP are effectively equivalent to microflows from a multipath load distribution standpoint. If the largest bandwidth requirement for any such LSP carried at that depth is known, then any link for which the "Maximum LSP Bandwidth" is less than that bandwidth requirement SHOULD be excluded from the path computation.

3.3.2. Link IP Capabilities and Path Computation

An MPLS-TP LSP cannot be reordered. There may be other types of LSP with strict packet ordering requirements. If LSP with strict packet ordering requirements carry IP, using IP headers in the multipath load distribution would violate the packet ordering requirements.

Some LSP cannot be reordered but do not carry IP, and do not carry payloads which could be mistaken as IP. For example, any LSP carrying only pseudowire traffic, where all pseudowires are using a control word carries no payloads which could be mistaken as IP. These type of LSP can be carried within MPLS LSP that allow use of IP header information in multipath load distribution.

3.3.2.1. LSP without Packet Ordering Requirements

Many LSP carry only IP or predominately IP, use no hierarchy or have little diversity in the MPLS label stack, and carry far more traffic than can be carried over a single component link in a multipath. Many LSP due to their high capacity, must traverse only multipath which will use IP header information in the multipath traffic distribution.

For these LSP, links must be excluded from the path computation which do not have the IPv4 Enabled Multipath and IPv6 Enabled Multipath bit set (if carrying both IPv4 and IPv6) and do not have either the Default to IP/MPLS Multipath bit set or the IP Optional Multipath bit set.

Hierarchical PSC LSP which require the use IP header information in the multipath traffic distribution MUST NOT set the Ordered Aggregate Enabled bit, MUST set the Default to IP/MPLS Multipath bit, and MUST NOT set the VARIP bit in the FA advertisement.

3.3.2.2. LSP with Ordering Requirements

In some cases an MPLS-TP may carry no IP traffic directly under the label stack. For example, if only pseudowire service ([RFC3985]) is being supported by an LSP, and all pseudowires are using a control word, and all control and management information is carried in a generic associated channel ([RFC5586]), then no IP traffic is carried directly under the label stack. In this case, it is highly desirable to signal the MPLS-TP LSP to allow IP header information to be used in the multipath load distribution. Doing so will allow any MPLS LSP containing this MPLS-TP LSP to allow the use of IP header information in the multipath load distribution.

Where MPLS-TP LSP are carrying IP, for any link for which the use of IP header information is not disabled or cannot be disabled on a per LSP basis, that link must be excluded from the path computation. Links which do not have to be excluded include link with the IPv4 Enabled Multipath and IPv6 Enabled Multipath bits clear, links with the Default to IP/MPLS Multipath clear, and links with the IP Optioal Multipath bit set. For those links with the IP Optioal Multipath set, MPLS-TP LSP which carry IP MUST explicitly disable the use of IP in the multipath load distribution in signaling if the Default to IP/MPLS Multipath is set and SHOULD explicitly disable the use o\ f IP in the multipath load distribution in signaling if the Default to IP/MPLS Multipath is clear.

3.3.2.3. LSP containing LSP with Ordering Requirements

The largest effective microflow with respect to a given multipath link can depend on whether the link can use IP header information in the multipath traffic distribution.

If a PSC LSP will need to carry traffic which cannot use IP header information in the multipath traffic distribution, then all links for which this capability is supported and enabled and cannot be disabled, must be excluded from the LSP path computation. Links which can be included in the LSP path computation include those with the IPv4 Enabled Multipath and IPv6 Enabled Multipath bits clear, those with the Default to IP/MPLS Multipath clear, or those with the VARIP set. For links with the IPv4 Enabled Multipath or IPv6 Enabled Multipath bit set, the Default to IP/MPLS Multipath bit set, and the VAR IP bit set, the requirement to disable use of IP in the multipath traffic distribution must be indicated in signaling.

3.3.3. Link Depth Limitations and Path Computation

For any LSP which does not have strict packet ordering constraints, LSP configuration SHOULD include the following parameters.

LSP Min Depth

a minimal acceptable number of label used in multipath traffic distribution,

LSP IP Depth

a minimal label stack depth where the IP header can be used in multipath traffic distribution

For example, if a PSC LSP will carry LSP which in turn carry very high capacity pseudowires using the pseudowire flow label (see [I-D.ietf-pwe3-fat-pw]), the flow label is four labels deep. In this case, LSP Min Depth should be configured at four or higher.

For example, if the same PSC LSP will carry LSP which carry IP traffic with no additional labels, then the IP header is two labels deep. In this case, LSP IP Depth should be configured at two or higher.

For an LSP with LSP Min Depth configured, all links with Max Depth set to a value below LSP Min Depth MUST be excluded from the LSP Path Computation.

For an LSP with LSP IP Depth configured, all links with IP Depth set to a value below LSP IP Depth MUST be excluded from the LSP Path Computation.

4. Backwards Compatibility

Networks today use three forms of multipath.

1. IP ECMP, including IP ECMP at LER using more than one LSP.
2. Ethernet Link Aggregation [IEEE-802.1AX].
3. MPLS Link Bundling [RFC4201].

4.1. Legacy Multipath Behavior

IP ECMP and Ethernet Link Aggregation always distribute traffic over the entire multipath either using information in the MPLS label stack, or using information in a potential IP header, or using both types of information.

One of two behaviors is assumed for link bundles. Either the link bundles place each LSP in its entirety on a single link bundle component link for all LSP, or link bundles distribute traffic over the entire link bundle using the same techniques used for ECMP and

Ethernet Link Aggregation. This second behavior is known as the "all ones" component link (see [RFC4201]).

4.2. Networks without Multipath Extensions

Networks exist that are comprised entirely of LSR which do not support these multipath extensions. In these networks there is no way of telling how multipath links will behave. Since an Ethernet Link Aggregation Group (LAG) is advertised as an ordinary link, there is no way to tell that it is a form of multipath.

4.2.1. Networks with MP Capability on all Multipath

Most large core network today rely heavily on the use of multipath. Ethernet Link Aggregation and configure LSR to use the "all ones" component link for all LSP. The "all ones" component link is the default for many Link Bundling implementations used in core networks.

This is equivalent to the following setting in the Multipath Node Capabilities sub-TLV or Multipath Link Capabilities sub-TLV.

Clear the Ordered Aggregate Enabled bit,

set the Multipath Enabled bit, set the Default to Multipath bit, and clearing the Variable Depth Multipath bit.

If the label stack is used in the multipath traffic distribution, set Max Depth to the number of label stack entries supported, otherwise set it to one.

If an IP packet under the label stack can be used in the multipath traffic distribution, set IPv4 Enabled Multipath, set IPv6 Enabled Multipath, set Default to IP/MPLS Multipath, and set IP Depth to the maximum number of label stack entries which can be skipped over before finding the IP stack. Otherwise clear IPv4 Enabled Multipath, clear IPv6 Enabled Multipath and clear Default to IP/MPLS Multipath.

These networks can support very large LSP but cannot support LSP which require strict packet ordering with other labels below such an LSP, such as pseudowire labels. They may also misroute OAM packet which use GAL (see [RFC5586]). Generally the Link Bundle advertisements indicate a "Maximum LSP Bandwidth" that is equal to the "Unreserved Bandwidth".

4.2.2. Networks with OA Capability on all Multipath

Some networks, particularly edge networks which tend to be lower capacity, do not use Link Aggregation, and if they use Link Bundling at all, configure each LSR to place each LSP in its entirety on a single link bundle component link for all LSP. Some edge equipment only supports this link bundle behavior.

This is equivalent to the following setting in the Multipath Node Capabilities sub-TLV or Multipath Link Capabilities sub-TLV.

Clear the Ordered Aggregate Enabled bit,

Clear the Multipath Enabled bit.

All remaining bits in the Flags field should be clear.

The Min Depth, Max Depth, and IP Depth should be set to zero.

These networks can support LSP which require strict packet ordering, but cannot support very large LSP.

4.2.3. Legacy Networks with Mixed MP and OA Links

Some network may support Ethernet Link Aggregation and all or a subset of LSR which place each LSP in its entirety on a single link bundle component link for all LSP.

If the "Maximum LSP Bandwidth" is set as described in Section 4.2.1, then very large LSP can be supported. Very large LSP cannot be supported on LSR which place each LSP in its entirety on a single link bundle component link for all LSP, but these are clearly indicated in signaling,

In these mixed networks it is not possible to reliably support LSP which require strict packet ordering. It is not possible to know where Ethernet Link Aggregation is used and it is not possible to accurately determine Link Bundling behavior on link bundles where "Maximum LSP Bandwidth" is equal to "Unreserved Bandwidth".

4.3. Transition to Multipath Extension Support

If a Multipath Node Capability sub-TLV is not advertised (see Section 2.1), then the LSR does not support these multipath extensions, or is not adjacent to any multipath. This allows detection of such nodes and if necessary application of defaults to cover Ethernet Link Aggregation Behavior.

4.3.1. Simple Transitions

For networks with LSR that do not support for multipath extensions, transition is easiest if all legacy LSR support and are configured with a common link bundling behavior. If Ethernet Link Aggregation is not used, a single configured default is needed to cover LSR that do not advertise a Multipath Node Capability sub-TLV.

If Ethernet Link Aggregation had been previously used on Legacy LSR, if possible LAG should be disabled and the members of the former LAG configured and advertised as a link bundle which uses the equivalent "all ones" behavior.

The transition network in this case lacks the ability to determine the largest microflow that can pass through legacy nodes, but this was the case prior to transition for the entire network prior to transition.

4.3.2. More Challenging Transitions

Transition is made more difficult if legacy LSR in a network support Ethernet Link Aggregation but do not support Link Bundle. This situation is most easily handled if a small upgrade to such an LSR can advertise a fixed Multipath Node Capability sub-TLV giving the characteristics of the Ethernet Link Aggregation on implementation on that node. Absent of such cooperation, the problem can be solved by configuration on newer LSR which allows association of a Multipath Node Capability sub-TLV with a specific legacy router ID and possibly a legacy router ID and link.

LSR supporting Multipath Extensions assign default values assigned by configuration to these legacy LSR running Ethernet Link Aggregation. These default values serve to allow LSP which require strict packet ordering to avoid these legacy LSR.

LSR which do not support [RFC4201] may be sufficiently rare that the ability to assign default values per legacy LSR may not be needed in practice.

5. IANA Considerations

[... to be completed ...]

The symbolic constants IANA-TBD-1 through IANA-TBD-4 need to be replaced. Complete instructions, including identification of the number space for each of these will be added to a later version of this internet-draft.

6. Security Considerations

The combination of MPLS, MPLS-TP, and multipath does not introduce any new security threats. The security considerations for MPLS/GMPLS and for MPLS-TP are documented in [RFC5920] and [I-D.ietf-mpls-tp-security-framework].

7. References

7.1. Normative References

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

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.
- [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 Packet Switched Network",
draft-ietf-pwe3-fat-pw-06 (work in progress), May 2011.
- [I-D.villamizar-mpls-tp-multipath]
Villamizar, C., "Use of Multipath with MPLS-TP and MPLS",
draft-villamizar-mpls-tp-multipath-01 (work in progress),
March 2011.
- [IEEE-802.1AX]
IEEE Standards Association, "IEEE Std 802.1AX-2008 IEEE Standard for Local and Metropolitan Area Networks - Link Aggregation", 2006, <<http://standards.ieee.org/getieee802/download/802.1AX-2008.pdf>>.
- [ITU-T.G.800]
ITU-T, "Unified functional architecture of transport networks", 2007, <<http://www.itu.int/rec/T-REC-G/recommendation.asp?parent=T-REC-G.800>>.
- [RFC2475] Blake, S., Black, D., Carlson, M., Davies, E., Wang, Z., and W. Weiss, "An Architecture for Differentiated Services", RFC 2475, December 1998.

- [RFC2702] Awduche, D., Malcolm, J., Agogbua, J., O'Dell, M., and J. McManus, "Requirements for Traffic Engineering Over MPLS", RFC 2702, September 1999.
- [RFC2991] Thaler, D. and C. Hopps, "Multipath Issues in Unicast and Multicast Next-Hop Selection", RFC 2991, November 2000.
- [RFC3260] Grossman, D., "New Terminology and Clarifications for Diffserv", RFC 3260, April 2002.
- [RFC3471] Berger, L., "Generalized Multi-Protocol Label Switching (GMPLS) Signaling Functional Description", RFC 3471, January 2003.
- [RFC3945] Mannie, E., "Generalized Multi-Protocol Label Switching (GMPLS) Architecture", RFC 3945, October 2004.
- [RFC3985] Bryant, S. and P. Pate, "Pseudo Wire Emulation Edge-to-Edge (PWE3) Architecture", RFC 3985, March 2005.
- [RFC4201] Kompella, K., Rekhter, Y., and L. Berger, "Link Bundling in MPLS Traffic Engineering (TE)", RFC 4201, October 2005.
- [RFC4206] Kompella, K. and Y. Rekhter, "Label Switched Paths (LSP) Hierarchy with Generalized Multi-Protocol Label Switching (GMPLS) Traffic Engineering (TE)", RFC 4206, October 2005.
- [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.
- [RFC5586] Bocci, M., Vigoureux, M., and S. Bryant, "MPLS Generic Associated Channel", RFC 5586, June 2009.
- [RFC5786] Aggarwal, R. and K. Kompella, "Advertising a Router's Local Addresses in OSPF Traffic Engineering (TE) Extensions", RFC 5786, March 2010.
- [RFC5920] Fang, L., "Security Framework for MPLS and GMPLS Networks", RFC 5920, July 2010.
- [RFC6107] Shiomoto, K. and A. Farrel, "Procedures for Dynamically Signaled Hierarchical Label Switched Paths", RFC 6107, February 2011.

Author's Address

Curtis Villamizar (editor)
Infinera Corporation
169 W. Java Drive
Sunnyvale, CA 94089

Email: curtis@occnc.com

Network Working Group
INTERNET-DRAFT
Intended Status: Standards Track
Expires: December 30, 2011

Sam Aldrin
Huawei Technologies
M.Venkatesan
Kannan KV Sampath
Aricent
Thomas D. Nadeau
CA Technologies
Sami Boutros
Cisco Systems
Ping Pan
Infinera

June 28, 2011

MPLS-TP Operations, Administration, and Management (OAM) Identifiers
Management Information Base (MIB)
draft-vkst-mpls-tp-oam-id-mib-00

Abstract

This memo defines a portion of the Management Information Base (MIB) for use with network management protocols in the Internet community. In particular, it describes Operations, Administration, and Management (OAM) identifiers related managed objects for Multiprotocol Label Switching (MPLS) based Transport Profile (TP).

Status of this Memo

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

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

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

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

The list of Internet-Draft Shadow Directories can be accessed at

<http://www.ietf.org/shadow.html>.

This Internet-Draft will expire on December 22, 2011.

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.

Table of Contents

1. Introduction	3
2. The Internet-Standard Management Framework	3
3. Overview	3
3.1 Conventions used in this document	3
3.2 Terminology	3
3.3 Acronyms	3
4. Feature List	4
5. Brief description of MIB Objects	4
5.1. mplsOamIdMegTable	4
5.2. mplsOamIdMeTable	4
6. Example of MPLS OAM identifier configuration for MPLS tunnel	5
7. MPLS OAM Identifiers MIB definitions	6
8. Security Consideration	21
9. IANA Considerations	22
10. References	22
10.1 Normative References	22
10.2 Informative References	22
11. Acknowledgments	23
12. Authors' Addresses	23

1 Introduction

This memo defines a portion of the Management Information Base (MIB) for use with network management protocols in the Internet community. In particular, it describes managed objects for modeling a Multiprotocol Label Switching (MPLS) [RFC3031] based transport profile.

This MIB module should be used for performing the OAM operations for MPLS LSPs and Pseudowires.

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 BCP 14, RFC2119.

2. The Internet-Standard Management Framework

For a detailed overview of the documents that describe the current Internet-Standard Management Framework, please refer to section 7 of RFC3410 [RFC3410].

Managed objects are accessed via a virtual information store, termed the Management Information Base or MIB. MIB objects are generally accessed through the Simple Network Management Protocol (SNMP). Objects in the MIB are defined using the mechanisms defined in the Structure of Management Information (SMI). This memo specifies a MIB module that is compliant to the SMIv2, which is described in STD 58, RFC2578, STD 58, RFC2579 and STD58, RFC2580.

3. Overview

3.1 Conventions used in this document

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

3.2 Terminology

This document uses terminology from the MPLS architecture document [RFC3031], MPLS Traffic Engineering Management information [RFC3812], MPLS Label Switch Router MIB [RFC3813] and MPLS-TP Identifiers document [TPIDS].

3.3 Acronyms

ICC: ITU Carrier Code
IP: Internet Protocol

LSP: Label Switching Path
LSR: Label Switching Router
MIB: Management Information Base
ME: Maintenance Entity
MEG: Maintenance Entity Group
MEP: Maintenance Entity Group End Point
MIP: Maintenance Intermediate Point
MPLS: Multi-Protocol Label Switching
MPLS-TP: MPLS Transport Profile
PW: Pseudowire
TE: Traffic Engineering
TP: Transport Profile

4. Feature List

The MPLS transport profile OAM identifiers MIB module is designed to satisfy the following requirements and constraints:

- The MIB module supports configuration of OAM identifiers for point-to-point, co-routed bi-directional, associated bi-directional MPLS tunnels and MPLS Pseudowires.

5. Brief description of MIB Objects

The objects described in this section support the functionality described in documents [RFC5654] and [TPIDS]. The tables support both IP compatible and ICC based OAM identifiers configurations for MPLS Tunnels and Pseudowires.

5.1. mplsOamIdMegTable

The mplsOamIdMegTable is used to manage one or more Maintenance Entities (MEs) that belongs to the same transport path.

When a new entry is created with mplsOamIdMegOperatorType set to ipCompatible (1), then as per [TPIDS] (MEG_ID for LSP is LSP_ID and MEG_ID for PW is PW_Path_ID), MEP_ID can be automatically formed.

For ICC based transport path, the user is expected to configure the ICC identifier explicitly in this table for MPLS tunnel and pseudowires.

5.2. mplsOamIdMeTable

The `mplsOamIdMeTable` defines a relationship between two points (source and sink) of a transport path to which maintenance and monitoring operations apply. The two points that define a maintenance entity are called Maintenance Entity Group End Points (MEPs).

In between MEPs, there are zero or more intermediate points, called Maintenance Entity Group Intermediate Points (MIPs). MEPs and MIPs are associated with the MEG and can be shared by more than one ME in a MEG.

6. Example of MPLS OAM identifier configuration for MPLS tunnel

In this section, we provide an example of the OAM identifier configuration for MPLS co-routed bidirectional tunnel.

This example provides usage of a MEG and ME tables for management and monitoring operations of MPLS tunnel.

This example considers the OAM identifiers configuration on a head-end LSR to manage and monitor a MPLS tunnel. Only relevant objects which are applicable for IP based OAM identifiers of co-routed MPLS tunnel are illustrated here.

In `mplsOamIdMegTable`:

```
{
  -- MEG index (Index to the table)
  mplsOamIdMegIndex          = 1,
  mplsOamIdMegName           = "MEG1",
  mplsOamIdMegOperatorType   = ipCompatible (1),
  mplsOamIdMegServiceType    = lsp (1),
  mplsOamIdMegMpLocation     = perNode(1),
  -- Mandatory parameters needed to activate the row go here
  mplsOamIdMegRowStatus      = createAndGo (4)
}
```

This will create an entry in the `mplsOamIdMegTable` to manage and monitor the MPLS tunnel.

The following ME table is used to associate the path information to a MEG.

In `mplsOamIdMeTable`:

```
{
  -- ME index (Index to the table)
  mplsOamIdMeIndex          = 1,
```

```

-- MP index (Index to the table)
mplsOamIdMeMpIndex          = 1,
mplsOamIdMeName              = "ME1",
mplsOamIdMeMpIfIndex        = 0,
-- Source MEP id is derived from the IP compatible MPLS tunnel
mplsOamIdMeSourceMepIndex    = 0,
-- Source MEP id is derived from the IP compatible MPLS tunnel
mplsOamIdMeSinkMepIndex      = 0,
mplsOamIdMeMpType            = mep (1),
mplsOamIdMeMepDirection     = down (2),
mplsOamIdMeProactiveOamPhbTCValue = 0,
mplsOamIdMeOnDemandOamPhbTCValue = 0,
-- RowPointer MUST point to the first accessible column of an
-- MPLS tunnel
mplsOamIdMeServicePointer    = mplsTunnelName.1.1.10.20,
-- Mandatory parameters needed to activate the row go here
mplsOamIdMeRowStatus         = createAndGo (4)
}

```

7. MPLS OAM Identifiers MIB definitions

```
MPLS-OAM-ID-STD-MIB DEFINITIONS ::= BEGIN
```

```

IMPORTS
    MODULE-IDENTITY, OBJECT-TYPE, NOTIFICATION-TYPE,
    Unsigned32, zeroDotZero
        FROM SNMPv2-SMI
        -- [RFC2578]
    MODULE-COMPLIANCE, OBJECT-GROUP, NOTIFICATION-GROUP
        FROM SNMPv2-CONF
        -- [RFC2580]
    RowStatus, TruthValue, RowPointer,
    DisplayString
        FROM SNMPv2-TC
        -- [RFC2579]
    mplsStdMIB
        FROM MPLS-TC-STD-MIB
        -- [RFC3811]
    InterfaceIndexOrZero, ifGeneralInformationGroup,
    ifCounterDiscontinuityGroup
        FROM IF-MIB;
        -- [RFC2863]

mplsOamIdStdMIB MODULE-IDENTITY
    LAST-UPDATED
        "201105300000Z" -- May 30, 2011
    ORGANIZATION
        "Multiprotocol Label Switching (MPLS) Working Group"
    CONTACT-INFO
        "
            Sam Aldrin
            Huawei Technologies, co.
            2330 Central Express Way,

```

Santa Clara, CA 95051, USA
Email: aldrin.ietf@gmail.com

Thomas D. Nadeau
CA Technologies
Email: thomas.nadeau@ca.com

Venkatesan Mahalingam
Aricent
India
Email: venkatesan.mahalingam@aricent.com

Kannan KV Sampath
Aricent
India
Email: Kannan.Sampath@aricent.com

Ping Pan
Infinera
Email: ppan@infinera.com

Sami Boutros
Cisco Systems, Inc.
3750 Cisco Way
San Jose, California 95134
USA
Email: sboutros@cisco.com

"

DESCRIPTION

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

This MIB module contains generic object definitions for
MPLS OAM maintenance identifiers in MPLS based transport
networks."

-- Revision history.

REVISION

"201105300000Z" -- May 30, 2011

DESCRIPTION

"MPLS OAM Identifiers mib objects for LSPs and
Pseudowires"

::= { mplsStdMIB xxx } -- xxx to be replaced with correct value

-- Top level components of this MIB module.

```
-- traps
mplsOamIdNotifications
    OBJECT IDENTIFIER ::= { mplsOamIdStdMIB 0 }
-- tables, scalars
mplsOamIdObjects OBJECT IDENTIFIER ::= { mplsOamIdStdMIB 1 }
-- conformance
mplsOamIdConformance
    OBJECT IDENTIFIER ::= { mplsOamIdStdMIB 2 }

-- Start of MPLS Transport Profile MEG table

mplsOamIdMegTable OBJECT-TYPE
    SYNTAX          SEQUENCE OF MplsTpOamIdMegEntry
    MAX-ACCESS      not-accessible
    STATUS          current
    DESCRIPTION
        "This table contains information about the Maintenance
        Entity Groups (MEG).

        MEG as mentioned in MPLS-TP OAM framework defines a set
        of one or more maintenance entities (ME).
        Maintenance Entities define a relationship between any
        two points of a transport path in an OAM domain to which
        maintenance and monitoring operations apply."
    ::= { mplsOamIdObjects 1 }

mplsOamIdMegEntry OBJECT-TYPE
    SYNTAX          MplsTpOamIdMegEntry
    MAX-ACCESS      not-accessible
    STATUS          current
    DESCRIPTION
        "An entry in this table represents MPLS-TP MEG.
        An entry can be created by a network administrator
        or by an SNMP agent as instructed by an MPLS-TP OAM
        Framework.

        When a new entry is created with
        mplsOamIdMegOperatorType set to ipCompatible (1),
        then as per [TPIDS] (MEG_ID for LSP is LSP_ID and
        MEG_ID for PW is PW_Path_ID), MEP_ID can be
        automatically formed.

        For LSP, MEP_ID is formed using
        Src-Global_ID::Src-Node_ID::Src-Tunnel_Num::LSP_Num.

        For PW, MEP_ID is formed using
        AGI::Src-Global_ID::Src-Node_ID::Src-AC_Id.MEG_ID.
```

MEP_ID is retrieved from the mplsOamIdMegServicePointer object based on the mplsOamIdMegServiceType value. ICC MEG_ID for LSP and PW is formed using the objects mplsOamIdMegIdIcc and mplsOamIdMegIdUmc.

MEP_ID can be formed using MEG_ID::MEP_Index."

REFERENCE

1. RFC 5860, Requirements for OAM in MPLS Transport Networks, May 2010.
2. MPLS-TP OAM Framework [TP-OAM-FWK].
3. MPLS-TP Identifiers [TPIDS]."

INDEX { mplsOamIdMegIndex }
 ::= { mplsOamIdMegTable 1 }

```
MplsTpOamIdMegEntry ::= SEQUENCE {
    mplsOamIdMegIndex      Unsigned32,
    mplsOamIdMegName       DisplayString,
    mplsOamIdMegOperatorType INTEGER,
    mplsOamIdMegIdIcc      DisplayString,
    mplsOamIdMegIdUmc      DisplayString,
    mplsOamIdMegServiceType INTEGER,
    mplsOamIdMegMpLocation INTEGER,
    mplsOamIdMegRowStatus  RowStatus
}
```

```
mplsOamIdMegIndex OBJECT-TYPE
    SYNTAX      Unsigned32
    MAX-ACCESS  not-accessible
    STATUS      current
    DESCRIPTION
        "Index for the conceptual row identifying a MEG within
         this MEG table."
    ::= { mplsOamIdMegEntry 1 }
```

```
mplsOamIdMegName OBJECT-TYPE
    SYNTAX      DisplayString (SIZE(1..48))
    MAX-ACCESS  read-create
    STATUS      current
    DESCRIPTION
        "Each Maintenance Entity Group has unique name amongst
         all those used or available to a service provider or
         operator. It facilitates easy identification of
         administrative responsibility for each MEG."
    ::= { mplsOamIdMegEntry 2 }
```

```
mplsOamIdMegOperatorType OBJECT-TYPE
    SYNTAX      INTEGER {
        ipCompatible (1),
```

```

        iccBased (2)
    }
MAX-ACCESS      read-create
STATUS          current
DESCRIPTION
    "Indicates the operator type for MEG. Conceptual rows
    having 'iccBased' as operator type, should have valid
    values for the objects mplsOamIdMegIdIcc and
    mplsOamIdMegIdUmc while making the row status active."
REFERENCE
    "MPLS-TP Identifiers [TPIDS]."
```

DEFVAL { ipCompatible }
 ::= { mplsOamIdMegEntry 3 }

mplsOamIdMegIdIcc OBJECT-TYPE
SYNTAX DisplayString (SIZE(1..6))
MAX-ACCESS read-write
STATUS current
DESCRIPTION
 "Unique code assigned to Network Operator or Service
 Provider maintained by ITU-T. The ITU Carrier Code
 used to form MEGID.

 This object contains non-null ICC value if
 the MplsTpOamIdMegOperatorType value is iccBased(2),
 otherwise null ICC value should be assigned."
REFERENCE
 "MPLS-TP Identifiers [TPIDS]."
DEFVAL { "" }
 ::= { mplsOamIdMegEntry 4 }

mplsOamIdMegIdUmc OBJECT-TYPE
SYNTAX DisplayString (SIZE(1..7))
MAX-ACCESS read-write
STATUS current
DESCRIPTION
 "Unique code assigned by Network Operator or Service
 Provider and is appended to mplsOamIdMegIdIcc to form
 the MEGID.

 This object contains non-null ICC value if
 the MplsTpOamIdMegOperatorType value is iccBased(2),
 otherwise null ICC value should be assigned."
REFERENCE
 "MPLS-TP Identifiers [TPIDS]."
DEFVAL { "" }
 ::= { mplsOamIdMegEntry 5 }

mplsOamIdMegServiceType OBJECT-TYPE

```
SYNTAX          INTEGER {
                lsp (1),
                pseudowire (2),
                section (3)
                }
MAX-ACCESS      read-create
STATUS          current
DESCRIPTION     "Indicates the service type for which the MEG is created.

                If the service type indicates lsp, the service pointer
                in mplsOamIdMeTable points to the TE tunnel table entry.

                If the value is pseudowire service type, the service
                pointer in mplsOamIdMeTable points to the pseudowire
                table entry.

                If the value is section service type, the service
                pointer in mplsOamIdMeTable points to a section entry."
DEFVAL { lsp }
 ::= { mplsOamIdMegEntry 6 }
```

mplsOamIdMegMpLocation OBJECT-TYPE

```
SYNTAX          INTEGER {
                perNode (1),
                perInterface (2)
                }
MAX-ACCESS      read-create
STATUS          current
DESCRIPTION     "Indicates the MP location type for this MEG.

                If the value is perNode, then the MEG in the LSR supports
                only perNode MEP/MIP, i.e., only one MEP/MIP in an LSR.

                If the value is perInterface, then the MEG in the LSR
                supports perInterface MEPs/MIPs, i.e., two MEPs/MIPs in
                an LSR."
REFERENCE      "MPLS-TP OAM draft, section 3.3 and 3.4"
DEFVAL { perNode }
 ::= { mplsOamIdMegEntry 7 }
```

mplsOamIdMegRowStatus OBJECT-TYPE

```
SYNTAX          RowStatus
MAX-ACCESS      read-create
STATUS          current
DESCRIPTION
```

```

    "This variable is used to create, modify, and/or delete
    a row in this table. When a row in this table is in
    active(1) state, no objects in that row can be modified
    by the agent except mplsOamIdMegRowStatus."
 ::= { mplsOamIdMegEntry 8 }

-- End of MPLS Transport Profile MEG table

-- Start of MPLS Transport Profile ME table
mplsOamIdMeTable OBJECT-TYPE
SYNTAX          SEQUENCE OF MplsTpOamIdMeEntry
MAX-ACCESS     not-accessible
STATUS         current
DESCRIPTION
    "This table contains MPLS-TP maintenance entity
    information.

    ME is some portion of a transport path that requires
    management bounded by two points (called MEPs), and the
    relationship between those points to which maintenance
    and monitoring operations apply.

    This table is generic enough to handle MEPs and MIPs
    information within a MEG."
 ::= { mplsOamIdObjects 2 }

mplsOamIdMeEntry OBJECT-TYPE
SYNTAX          MplsTpOamIdMeEntry
MAX-ACCESS     not-accessible
STATUS         current
DESCRIPTION
    "An entry in this table represents MPLS-TP maintenance
    entity. This entry represents the ME if the source and
    sink MEPs are defined.

    A ME is a p2p entity. One ME has two such MEPs.
    A MEG is a group of one or more MEs. One MEG can have
    two or more MEPs.

    For P2P LSP, one MEG has one ME and this ME is associated
    two MEPs (source and sink MEPs) within a MEG.
    Each mplsOamIdMeIndex value denotes the ME within a MEG.

    In case of unidirectional point-to-point transport paths,
    a single unidirectional Maintenance Entity is defined to
    monitor it.

    In case of associated bi-directional point-to-point
```

transport paths, two independent unidirectional Maintenance Entities are defined to independently monitor each direction. This has implications for transactions that terminate at or query a MIP, as a return path from MIP to source MEP does not necessarily exist within the MEG.

In case of co-routed bi-directional point-to-point transport paths, a single bidirectional Maintenance Entity is defined to monitor both directions congruently.

In case of unidirectional point-to-multipoint transport paths, a single unidirectional Maintenance entity for each leaf is defined to monitor the transport path from the root to that leaf."

```
INDEX { mplsOamIdMegIndex,
        mplsOamIdMeIndex,
        mplsOamIdMeMpIndex
      }
 ::= { mplsOamIdMeTable 1 }
```

```
MplsTpOamIdMeEntry ::= SEQUENCE {
    mplsOamIdMeIndex           Unsigned32,
    mplsOamIdMeMpIndex        Unsigned32,
    mplsOamIdMeName           DisplayString,
    mplsOamIdMeMpIfIndex      InterfaceIndexOrZero,
    mplsOamIdMeSourceMepIndex Unsigned32,
    mplsOamIdMeSinkMepIndex   Unsigned32,
    mplsOamIdMeMpType         INTEGER,
    mplsOamIdMeMepDirection   INTEGER,
    mplsOamIdMeProactiveOamSessIndex Unsigned32,
    mplsOamIdMeProactiveOamPhbTCValue INTEGER,
    mplsOamIdMeOnDemandOamPhbTCValue INTEGER,
    mplsOamIdMeServiceSignaled TruthValue,
    mplsOamIdMeServicePointer RowPointer,
    mplsOamIdMeRowStatus      RowStatus
  }
```

```
mplsOamIdMeIndex OBJECT-TYPE
    SYNTAX          Unsigned32
    MAX-ACCESS      not-accessible
    STATUS          current
    DESCRIPTION
        "Uniquely identifies a maintenance entity index within
         a MEG."
    ::= { mplsOamIdMeEntry 1 }
```

```
mplsOamIdMeMpIndex OBJECT-TYPE
```

```
SYNTAX          Unsigned32
MAX-ACCESS      not-accessible
STATUS          current
DESCRIPTION
  "Indicates the maintenance point index.
  The value of this object can be MEP index or MIP index."
 ::= { mplsOamIdMeEntry 2 }

mplsOamIdMeName OBJECT-TYPE
SYNTAX          DisplayString (SIZE(1..48))
MAX-ACCESS      read-create
STATUS          current
DESCRIPTION
  "This object denotes the ME name, each
  Maintenance Entity has unique name within MEG."
 ::= { mplsOamIdMeEntry 3 }

mplsOamIdMeMpIfIndex OBJECT-TYPE
SYNTAX          InterfaceIndexOrZero
MAX-ACCESS      read-create
STATUS          current
DESCRIPTION
  "Indicates the maintenance point interface.
  If the mplsOamIdMegMpLocation object value
  is perNode (1), the MP interface index should point
  to incoming interface or outgoing interface or
  zero (indicates the MP OAM packets are initiated
  from forwarding engine).

  If the mplsOamIdMegMpLocation object value is
  perInterface (2), the MP interface index should point to
  incoming interface or outgoing interface."
REFERENCE
  "MPLS-TP OAM framework draft, 3.3 and 3.4"
DEFVAL { 0 }
 ::= { mplsOamIdMeEntry 4 }

mplsOamIdMeSourceMepIndex OBJECT-TYPE
SYNTAX          Unsigned32
MAX-ACCESS      read-create
STATUS          current
DESCRIPTION
  "Indicates the source MEP Index of the ME. This object
  should be configured if mplsOamIdMegOperatorType object
  in the mplsOamIdMegEntry is configured as iccBased (2).
  If the MEG is configured for IP based operator,
  the value of this object should be set zero and the MEP
  ID will be automatically derived from the service
```

```

        Identifiers(MPLS-TP LSP/PW Identifier)."  

    DEFVAL { 0 }  

    ::= { mplsOamIdMeEntry 5 }  
  

mplsOamIdMeSinkMepIndex OBJECT-TYPE  

    SYNTAX      Unsigned32  

    MAX-ACCESS  read-create  

    STATUS      current  

    DESCRIPTION  

        "Indicates the sink MEP Index of the ME. This object  

        should be configured if mplsOamIdMegOperatorType object  

        in the mplsOamIdMegEntry is configured as iccBased (2).  

        If the MEG is configured for IP based operator,  

        the value of this object should be set zero and the MEP  

        ID will be automatically derived from the service  

        Identifiers(MPLS-TP LSP/PW Identifier)."  

    DEFVAL { 0 }  

    ::= { mplsOamIdMeEntry 6 }  
  

mplsOamIdMeMpType OBJECT-TYPE  

    SYNTAX      INTEGER {  

                mep (1),  

                mip (2)  

            }  

    MAX-ACCESS  read-create  

    STATUS      current  

    DESCRIPTION  

        "Indicates the maintenance point type within the MEG.  
  

        The object should have the value mep (1), only in the  

        Ingress or Egress nodes of the transport path.  
  

        The object can have the value mip (2),  

        in the intermediate nodes and possibly in the end nodes  

        of the transport path."  

    DEFVAL { mep }  

    ::= { mplsOamIdMeEntry 7 }  
  

mplsOamIdMeMepDirection OBJECT-TYPE  

    SYNTAX      INTEGER {  

                up (1),  

                down (2)  

            }  

    MAX-ACCESS  read-create  

    STATUS      current  

    DESCRIPTION  

        "Indicates the direction of the MEP. This object  

        should be configured if mplsOamIdMeMpType is
```

```
        configured as mep (1)."  
    DEFVAL { down }  
    ::= { mplsOamIdMeEntry 8 }  
  
mplsOamIdMeProactiveOamSessIndex OBJECT-TYPE  
    SYNTAX      Unsigned32  
    MAX-ACCESS  read-only  
    STATUS      current  
    DESCRIPTION  
        "Indicates the proactive OAM session index for this MP.  
        When a proactive OAM session for this MP is established,  
        the underlying proactive initiator has to update this  
        object with the proactive OAM session index."  
    DEFVAL { 0 }  
    ::= { mplsOamIdMeEntry 9 }  
  
mplsOamIdMeProactiveOamPhbTCValue OBJECT-TYPE  
    SYNTAX      INTEGER {  
                efl (1),  
                efl2 (2),  
                afl (3),  
                afl2 (4),  
                afl3 (5),  
                be (6)  
            }  
    MAX-ACCESS  read-create  
    STATUS      current  
    DESCRIPTION  
        "Indicates the Per-hop Behavior (PHB) value for this source  
        MEP generated proactive traffic."  
    DEFVAL { efl }  
    ::= { mplsOamIdMeEntry 10 }  
  
mplsOamIdMeOnDemandOamPhbTCValue OBJECT-TYPE  
    SYNTAX      INTEGER {  
                efl (1),  
                efl2 (2),  
                afl (3),  
                afl2 (4),  
                afl3 (5),  
                be (6)  
            }  
    MAX-ACCESS  read-create  
    STATUS      current  
    DESCRIPTION  
        "Indicates the Per-hop Behavior (PHB) value for this  
        source MEP generated on-demand traffic."  
    DEFVAL { efl }
```

```
 ::= { mplsOamIdMeEntry 11 }

mplsOamIdMeServiceSignaled OBJECT-TYPE
    SYNTAX          TruthValue
    MAX-ACCESS      read-create
    STATUS          current
    DESCRIPTION
        "Indicates whether the service associated with ME is
         created by signaling or static."
    DEFVAL { false }
 ::= { mplsOamIdMeEntry 12 }

mplsOamIdMeServicePointer OBJECT-TYPE
    SYNTAX          RowPointer
    MAX-ACCESS      read-create
    STATUS          current
    DESCRIPTION
        "This variable represents a pointer to the MPLS-TP
         transport path. This value may point at an entry in the
         mplsTunnelEntry if mplsOamIdMegServiceType is configured
         as lsp (1) or at an entry in the pwEntry if
         mplsOamIdMegServiceType is configured as pseudowire (2).

         Note: This service pointer object, is placed in ME table
         instead of MEG table, since it will be useful in case of
         point-to-multipoint, where each ME will point to different
         branches of a P2MP tree."
    DEFVAL { zeroDotZero }
 ::= { mplsOamIdMeEntry 13 }

mplsOamIdMeRowStatus OBJECT-TYPE
    SYNTAX          RowStatus
    MAX-ACCESS      read-create
    STATUS          current
    DESCRIPTION
        "This variable is used to create, modify, and/or
         delete a row in this table. When a row in this
         table is in active(1) state, no objects in that row
         can be modified by the agent except
         mplsOamIdMeRowStatus."
 ::= { mplsOamIdMeEntry 14 }

-- End of MPLS Transport Profile ME table

-- End of MPLS-TP OAM Tables

-- Trap Definitions of MPLS-TP identifiers
```

```
mplsOamIdMegOperStatus OBJECT-TYPE
    SYNTAX      INTEGER {
                up (1),
                down (2)
                }
    MAX-ACCESS  accessible-for-notify
    STATUS      current
    DESCRIPTION
        "This object specifies the operational status of the
        Maintenance Entity Group (MEG). This object is used to
        send the notification to the SNMP manager about the MEG.

        The value up (1) indicates that the MEG and its monitored
        path are operationally up. The value down (2) indicates
        that the MEG is operationally down."
 ::= { mplsOamIdObjects 3 }

mplsOamIdMegSubOperStatus OBJECT-TYPE
    SYNTAX      BITS {
                megDown (0),
                meDown (1),
                oamAppDown (2),
                pathDown (3)
                }
    MAX-ACCESS  accessible-for-notify
    STATUS      current
    DESCRIPTION
        "This object specifies the reason why the MEG operational
        status as mentioned by the object mplsOamIdMegOperStatus
        is down. This object is used to send the notification to
        the SNMP manager about the MEG.

        The bit 0 (megDown) when set indicates the MEG is down.
        MEG table can be made down administratively.
        The bit 1 (meDown) when set indicates the ME table is
        down. ME can be made down administratively.
        The bit 2 (oamAppDown) when set indicates that the
        OAM application has notified that the entity (LSP or PW)
        monitored by this MEG is down. Currently, BFD is the
        only supported OAM application.
        The bit 3 (pathDown) when set indicates that the
        underlying LSP or PW is down."
 ::= { mplsOamIdObjects 4 }

mplsOamIdDefectCondition NOTIFICATION-TYPE
    OBJECTS      {
                mplsOamIdMegName,
                mplsOamIdMeName,
```

```

        mplsOamIdMegOperStatus,
        mplsOamIdMegSubOperStatus
    }
STATUS      current
DESCRIPTION
    "This notification signifies the operational status of MEG.

    The information that are carried in this notification are
    Meg Name, Me Name, MegOperStatus and
    MegSubOperStatus.
    "
 ::= { mplsOamIdNotifications 1 }

-- End of Notifications.

-- Module Compliance.

mplsOamIdGroups
    OBJECT IDENTIFIER ::= { mplsOamIdConformance 1 }

mplsOamIdCompliances
    OBJECT IDENTIFIER ::= { mplsOamIdConformance 2 }

-- Compliance requirement for fully compliant implementations.

mplsOamIdModuleFullCompliance MODULE-COMPLIANCE
STATUS      current
DESCRIPTION "Compliance statement for agents that provide full
            support for MPLS-TP-OAM-STD-MIB. Such devices can
            then be monitored and also be configured using
            this MIB module."

MODULE IF-MIB -- The Interfaces Group MIB, RFC 2863.
MANDATORY-GROUPS {
    ifGeneralInformationGroup,
    ifCounterDiscontinuityGroup
}

MODULE -- This module.
MANDATORY-GROUPS {
    mplsOamIdMegGroup,
    mplsOamIdMeGroup
}

GROUP      mplsOamIdTrapGroup
DESCRIPTION "This group is only mandatory for those
            implementations which can efficiently implement
            the notifications contained in this group."
```

```
GROUP          mplsOamIdNotificationGroup
DESCRIPTION   "This group is only mandatory for those
               implementations which can efficiently implement
               the notifications contained in this group."

 ::= { mplsOamIdCompliances 1 }

-- Units of conformance.

mplsOamIdMegGroup OBJECT-GROUP
OBJECTS {
    mplsOamIdMegName,
    mplsOamIdMegOperatorType,
    mplsOamIdMegIdIcc,
    mplsOamIdMegIdUmc,
    mplsOamIdMegServiceType,
    mplsOamIdMegMpLocation,
    mplsOamIdMegRowStatus
}

STATUS current
DESCRIPTION
    "Collection of objects needed for MPLS MEG information."
 ::= { mplsOamIdGroups 1 }

mplsOamIdMeGroup OBJECT-GROUP
OBJECTS {
    mplsOamIdMeName,
    mplsOamIdMeMpIfIndex,
    mplsOamIdMeSourceMepIndex,
    mplsOamIdMeSinkMepIndex,
    mplsOamIdMeMpType,
    mplsOamIdMeMepDirection,
    mplsOamIdMeProactiveOamSessIndex,
    mplsOamIdMeProactiveOamPhbTCValue,
    mplsOamIdMeOnDemandOamPhbTCValue,
    mplsOamIdMeServiceSignaled,
    mplsOamIdMeServicePointer,
    mplsOamIdMeRowStatus
}
STATUS current
DESCRIPTION
    "Collection of objects needed for MPLS ME information."
 ::= { mplsOamIdGroups 2 }

mplsOamIdTrapGroup OBJECT-GROUP
OBJECTS {
    mplsOamIdMegOperStatus,
```

```

    mplsOamIdMegSubOperStatus
  }
  STATUS current
  DESCRIPTION
    "Collection of objects needed to implement notifications."
  ::= { mplsOamIdGroups 3 }

mplsOamIdNotificationGroup NOTIFICATION-GROUP
  NOTIFICATIONS {
    mplsOamIdDefectCondition
  }
  STATUS current
  DESCRIPTION
    "Set of notifications implemented in this module."
  ::= { mplsOamIdGroups 4 }

END

```

8. Security Consideration

There is a number of management objects defined in this MIB module that has a MAX-ACCESS clause of read-write.. Such objects may be considered sensitive or vulnerable in some network environments. The support for SET operations in a non-secure environment without proper protection can have a negative effect on network operations.

Some of the readable objects in this MIB module (i.e., objects with a MAX-ACCESS other than not-accessible) may be considered sensitive or vulnerable in some network environments. It is thus important to control even GET and/or NOTIFY access to these objects and possibly to even encrypt the values of these objects when sending them over the network via SNMP. These are the tables and objects and their sensitivity/vulnerability:

SNMP versions prior to SNMPv3 did not include adequate security. Even if the network itself is secure (for example by using IPsec), even then, there is no control as to who on the secure network is allowed to access and GET/SET (read/change/create/delete) the objects in this MIB module.

It is RECOMMENDED that implementers consider the security features as provided by the SNMPv3 framework (see [RFC3410], section 8), including full supports for the SNMPv3 cryptographic mechanisms (for authentication and privacy).

Further, deployment of SNMP versions prior to SNMPv3 is NOT RECOMMENDED. Instead, it is RECOMMENDED to deploy SNMPv3 and to

enable cryptographic security. It is then a customer/operator responsibility to ensure that the SNMP entity giving access to an instance of this MIB module is properly configured to give access to the objects only to those principles (users) that have legitimate rights to indeed GET or SET (change/create/delete) them.

9. IANA Considerations

To be added in a later version of this document.

10. References

10.1 Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [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, January 2001.

10.2 Informative References

- [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.
- [RFC3410] J. Case, R. Mundy, D. pertain, B.Stewart, "Introduction and Applicability Statement for Internet Standard

Management Framework", RFC 3410, December 2002.

[RFC3811] Nadeau, T., Ed., and J. Cucchiara, Ed., "Definitions of Textual Conventions (TCs) for Multiprotocol Label Switching (MPLS) Management", RFC 3811, June 2004.

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

[TPIDS] M. Bocci, et al, "MPLS-TP Identifiers", draft-ietf-mpls-tp-identifiers-03, October 25, 2010

[TP-OAM-FWK]
Bocci, M. and D. Allan, "Operations, Administration and Maintenance Framework for MPLS-based Transport Networks", 2010, <draft-ietf-mpls-tp-oam-framework-10.txt>.

11. Acknowledgments

To be added in a later version of this document.

12. Authors' Addresses

Sam Aldrin
Huawei Technologies, co.
2330 Central Express Way,
Santa Clara, CA 95051, USA
Email: aldrin.ietf@gmail.com

Thomas D. Nadeau
CA Technologies,
Email: thomas.nadeau@ca.com

Venkatesan Mahalingam
Aricent
India
Email: venkatesan.mahalingam@aricent.com

Kannan KV Sampath
Aricent
India
Email: Kannan.Sampath@aricent.com

Ping Pan
Infinera
Email: ppan@infinera.com

Sami Boutros
Cisco Systems, Inc.
3750 Cisco Way
San Jose, California 95134
USA
Email: sboutros@cisco.com

Internet Engineering Task Force
Internet-Draft
Intended status: Standards Track
Expires: January 9, 2012

Q. Zhao
Huawei Technology
L. Fang
C. Zhou
Cisco Systems
L. Li
ChinaMobile
N. So
Verison Business
R. Torvi
Juniper Networks
July 8, 2011

LDP Extension for Multi Topology Support
draft-zhao-mpls-ldp-multi-topology-02.txt

Abstract

Multi-Topology (MT) routing is supported in IP through extension of IGP protocols, such as OSPF and IS-IS. Each route computed by OSPF or IS-IS is associated with a specific topology. Label Distribution Protocol (LDP) is used to distribute labels for FECs advertised by routing protocols. It is a natural requirement to extend LDP in order to make LDP be aware of MT and thus take advantage of MT based routing.

This document describes options to extend the existing MPLS signalling protocol (LDP) for creating and maintaining Label Switching Paths (LSPs) in a Multi-Topology enviroment.

Status of this Memo

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

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

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

material or to cite them other than as "work in progress."

This Internet-Draft will expire on January 9, 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. Terminology	4
2. Introduction	4
3. Application Scenarios	5
3.1. Simplified Data-plane	5
3.2. Using MT for p2p Protection	6
3.3. Using MT for mLDP Protection	6
3.4. Service Separation	6
3.5. Simplified inter-AS VPN Solution	6
4. Associating a FEC or group of FECs with MT-ID	7
4.1. MT-ID TLV	7
4.2. FEC TLV with MT-ID Extension	8
5. LDP MT Capability Advertisement	9
5.1. Session Initialization	10
5.2. After Session Setup	11
6. LDP Sessions	12
7. Reserved MT ID Values	12
8. LDP Messages with FEC TLV and MT-ID TLV	12
8.1. Label Mapping Message	13
8.2. Label Request Message	14
8.3. Label Abort Request Message	14
8.4. Label Withdraw Message	15
8.5. Label Release Message	16
9. Session Initialization Message with MT Capability	16
10. MPLS Forwarding in MT	17
10.1. Use Label for (FEC, MT-ID) Tuple	17
11. Security Consideration	18
12. IANA Considerations	18
13. Acknowledgement	18
14. References	19
14.1. Normative References	19
14.2. Informative References	19
Authors' Addresses	19

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

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

OSPF and IS-IS use MT-ID (Multi-Topology Identification) to identify different topologies. For each topology identified by a MT-ID, IGP computes a separate SPF tree independently to find the best paths to the IP prefixes associated with this topology.

For FECs that are associated with a specific topology, we propose to use the same MT-ID of this topology in LDP. Thus the Label Switching Path (LSP) for a certain FEC may be created and maintained along the IGP path in this topology.

Maintaining multiple MTs for MPLS network in a backwards-compatible manner requires several extensions to the label signaling encoding and processing procedures. When label is associated with a FEC, the FEC includes both ip address and topology it belongs to.

There are two possible solutions to support MT aware MPLS network from MPLS forwarding point of view. The first one is to map label to both ip address and the corresponding topology. The alternative one is to use label stacks. The upper label maps to the topology, the lower label maps to the ip address. The first option does not require change to data plane, and it could use multiple labels for the same address on different topologies. The second option requires two lookups on data forwarding plane, and it can use the same label

for the same address on different topologies.

There are a few possible ways to apply the MT-ID of a topology in LDP. One way is to have a new TLV for MT-ID and insert the TLV into messages describing a FEC that needs Multi-Topology information. Another approach is to expand the FEC TLV to contain MT-ID if the FEC needs Multi-Topology information.

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 landlines 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 map an attribute of incoming IP packet header to IGP-MT, which is additional

overhead for network management. Whereas, MPLS-MT mapping is required only at the ingress-PE of an MPLS-MT LSP, because of each node identifies MPLS-MT LSP switching based on incoming label, hence no additional configuration is required at every node.

3.2. Using MT for p2p Protection

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

3.3. Using MT for mLDP Protection

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

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 to separate multicast and unicast traffic.

3.5. Simplified inter-AS VPN Solution

When the LSP is crossing multiple domains for the inter-AS VPN scenarios, the LSP setup process can be simplified by configuring a

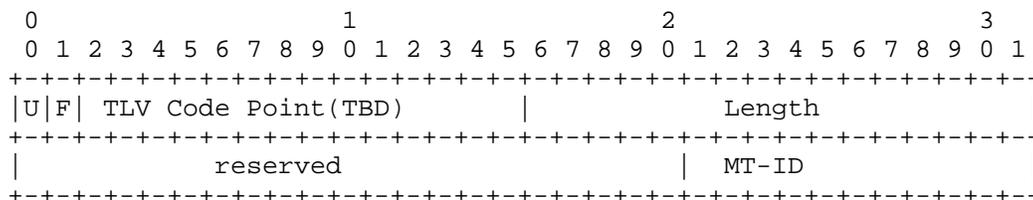
set of routers which are in different domains into a new single domain with a new topology ID using the LDP multiple topology. All the routers belong this new topology will be used to carry the traffic acrossing multiple domains and since they are in a single domain with the new topology ID, so the LDP 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 FEC or group of FECs with MT-ID

This section describes two approaches to associate a FEC or a group of FECs to a MT-ID in LDP. One way is to have a new TLV for MT-ID and insert the MT-ID TLV into messages describing a FEC that needs Multi-Topology information. Another approach is to extend FEC TLV to contain the MT-ID if the FEC needs Multi-Topology information.

4.1. MT-ID TLV

The new TLV for MT-ID is defined as below:



where:

U and F bits:
As specified in [RFC3036].

TLV Code Point:
The TLV type which identifies a specific capability.

MT-ID is a 12-bit field containing the ID of the topology corresponding to the MT-ID used in IGP and LDP. Lack of MT-ID TLV in messages MUST be interpreted as FECs are used in default MT-ID (0) only.

A MT-ID TLV can be inserted into the following LDP messages as an optional parameter.

Label Mapping	"Label Mapping Message"
Label Request	"Label Request Message"
Label Abort Request	"Label Abort Request Message"
Label Withdraw	"Label Withdraw Message"
Label Release	"Label Release Message"

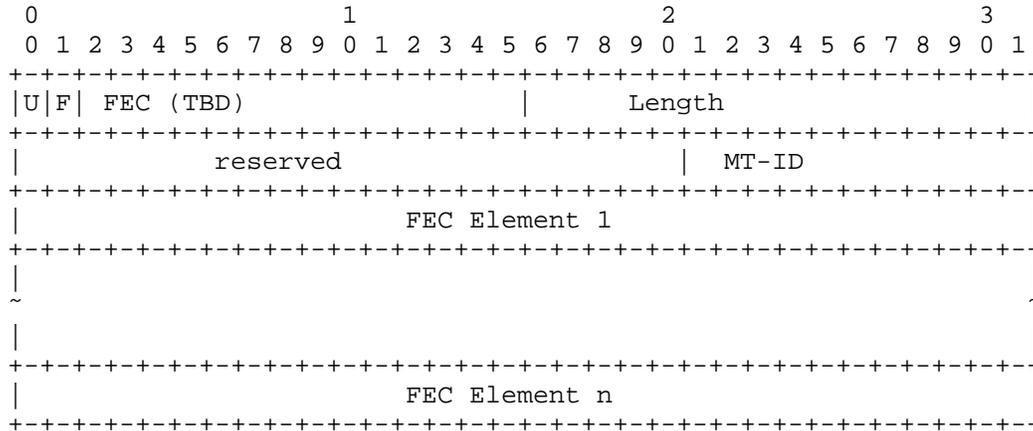
The message with inserted MT-ID TLV associates a FEC in same message to the topology identified by MT-ID.

Figure 1: MT-ID TLV Format

4.2. FEC TLV with MT-ID Extension

The new TLV for MT-ID is defined as below:

The extended FEC TLV has the format below.



This new FEC TLV may contain a number of FEC elements and a MT-ID. It associates these FEC elements with the topology identified by the MT-ID. Each FEC TLV can contain only one MT-ID.

Figure 2: Extended FEC with MT-ID

5. LDP MT Capability Advertisement

The LDP MT capability can be advertised either during the LDP session initialization or after the LDP session is setup.

The capability for supporting multi-topology in LDP can be advertised during LDP session initialization stage by including the LDP MT capability TLV in LDP Initialization message. After LDP session is established, the MT capability can also be advertised or changed using Capability message.

If an LSR has not advertised MT capability, its peer must not send messages that include MT identifier to this LSR.

If an LSR receives a Label Mapping message with MT parameter from downstream LSR-D and its upstream LSR-U has not advertised MT capability, an LSP for the MT will not be established.

If an LSR is changed from non MT capable to MT capable, it sets the S bit in MT capability TLV and advertises via the Capability message. The existing LSP is treated as LSP for default MT (ID 0).

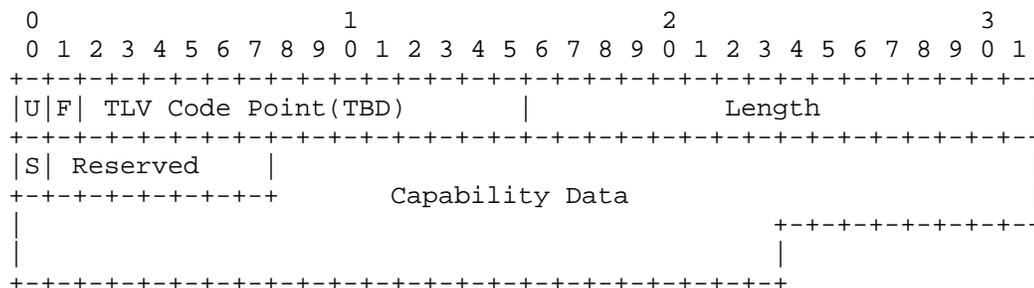
If an LSR is changed from MT capable to non-MT capable, it may initiate withdraw of all label mapping for existing LSPs of all non-default MTs. Alternatively, it may wait until the routing update to withdraw FEC and release the label mapping for existing LSPs of specific MT.

There will be case where IGP is MT capable but MPLS is not and the handling procedure for this case is TBD.

5.1. Session Initialization

In an LDP session initialization, the MT capability may be advertised through an extended session initialization message. This extended message has the same format as the original session initialization message but contains the LDP MT capability TLV as an optional parameter.

The format of the TLV for LDP MT is specified in the [LDPCAP] as below:



where:

U and F bits:
As specified in [RFC3036].

TLV Code Point:
The TLV type which identifies a specific capability. The "IANA Considerations" section of [RFC3036] specifies the assignment of code points for LDP TLVs.

S-bit:
The State Bit indicates whether the sender is advertising or withdrawing the capability corresponding to the TLV Code Point. The State bit is used as follows:

- 1 - The TLV is advertising the capability specified by the TLV Code Point.
- 0 - The TLV is withdrawing the capability specified by the TLV Code Point.

Capability Data:

Information, if any, about the capability in addition to the TLV Code Point required to fully specify the capability.

Figure 3: LDP MT CAP TLV

5.2. After Session Setup

During the normal operating stage of LDP sessions, the capability message defined in the [LDPCAP] will be used with an LDP MT capability TLV.

The format of the Capability message is as follows:

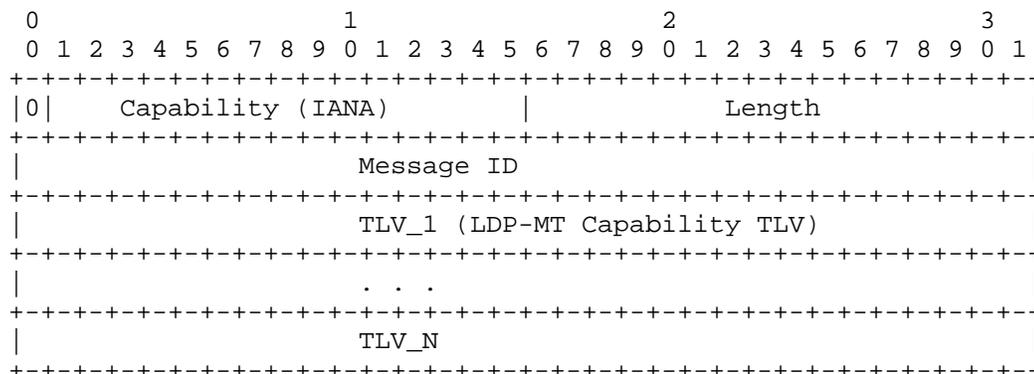


Figure 4: LDP CAP Format

where TLV_1 (LDP-MT Capability TLV) specifies that the LDP MT capability is enabled or disabled by setting the S bit of the TLV to 1 or 0.

6. LDP Sessions

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

This section is still TBD.

7. Reserved MT ID Values

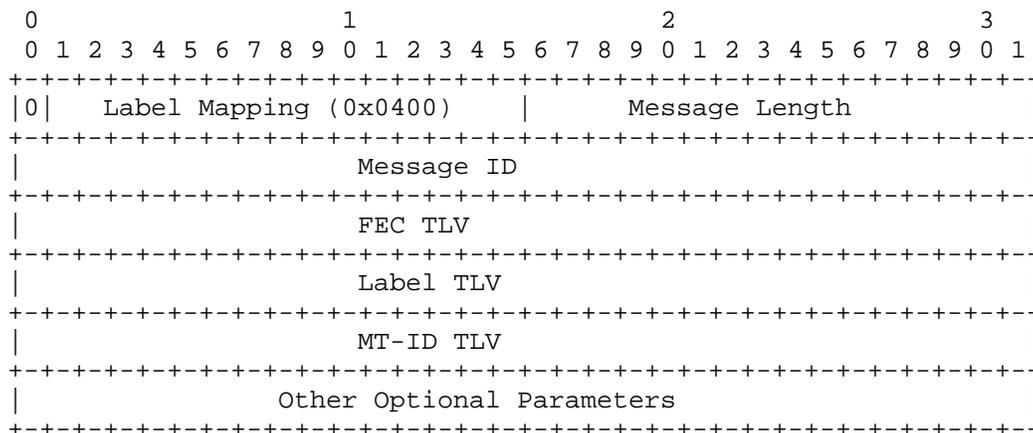
Certain MT topologies are assigned to serve pre-determined purposes:
[TBD]

8. LDP Messages with FEC TLV and MT-ID TLV

8.1. Label Mapping Message

An LSR sends a Label Mapping message to an LDP peer to advertise FEC-label bindings. In the Optional Parameters' field, the MT-ID TLV will be inserted.

The encoding for the Label Mapping message is:



Optional Parameters

This variable length field contains 0 or more parameters, each encoded as a TLV. The optional parameters are:

Optional Parameter	Length	Value
Label Request	4	See below
Message ID TLV		
Hop Count TLV	1	See below
Path Vector TLV	variable	See below
MT TLV	variable	See below

MT TLV

see the defination section for this new TLV.

Figure 5: Label Mapping Message

8.2. Label Request Message

An LSR sends the Label Request message to an LDP peer to request a binding (mapping) for a FEC. The MT TLV will be inserted into the Optional parameters' field.

The encoding for the Label Request message is:

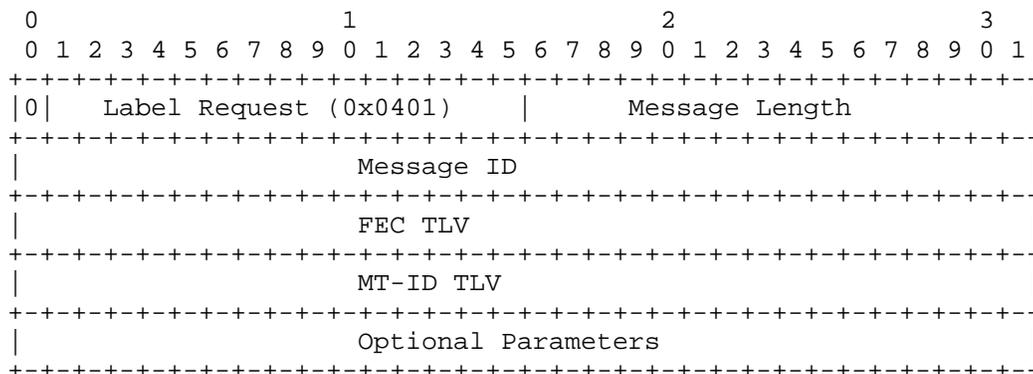


Figure 6: Label Request Message

In the DU mode, when a label mapping is received by a LSR which has a downstream with MT capability advertised and an upstream without the MT capability advertised, it will not send label mapping to its upstream.

in the DoD mode, the label request is sent down to the downstream LSR until it finds the downstream LSR which doesn't support the MT, then the current LSPR will send a notification to its upstream LSR. In this case, no LSP is setup.

We propose to add a new notification event to signal the upstream that the downstream is not capable.

8.3. Label Abort Request Message

The Label Abort Request message may be used to abort an outstanding Label Request message. The MT TLV may be inserted into the optional parameters' field.

The encoding for the Label Abort Request message is:

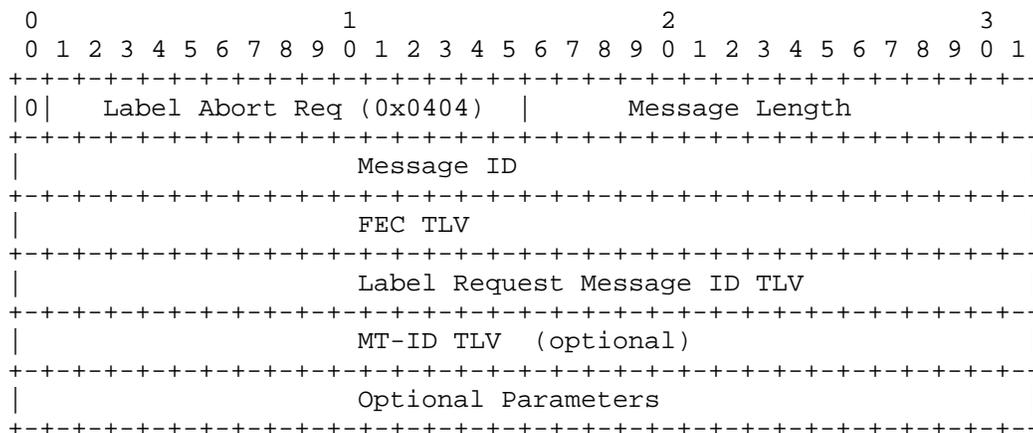


Figure 7: Label Abort Request Message

8.4. Label Withdraw Message

An LSR sends a Label Withdraw Message to an LDP peer to signal the peer that the peer may not continue to use specific FEC-label mappings the LSR had previously advertised. This breaks the mapping between the FECs and the labels. The MT TLV will be added into the optional parameters' field.

The encoding for the Label Withdraw Message is:

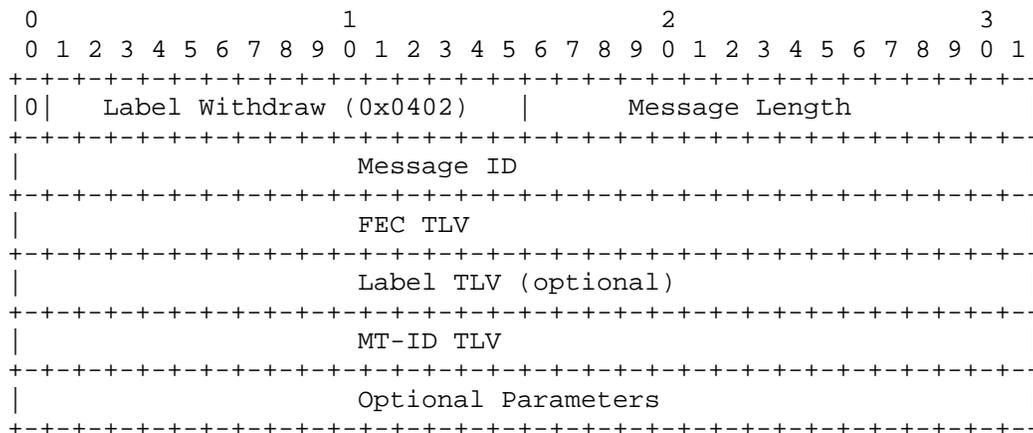


Figure 8: Label Withdraw Message

8.5. Label Release Message

An LSR sends a Label Release message to an LDP peer to signal the peer that the LSR no longer needs specific FEC-label mappings previously requested of and/or advertised by the peer. The MT TLV will be added into the optional parameters' field.

The encoding for the Label Release Message is:

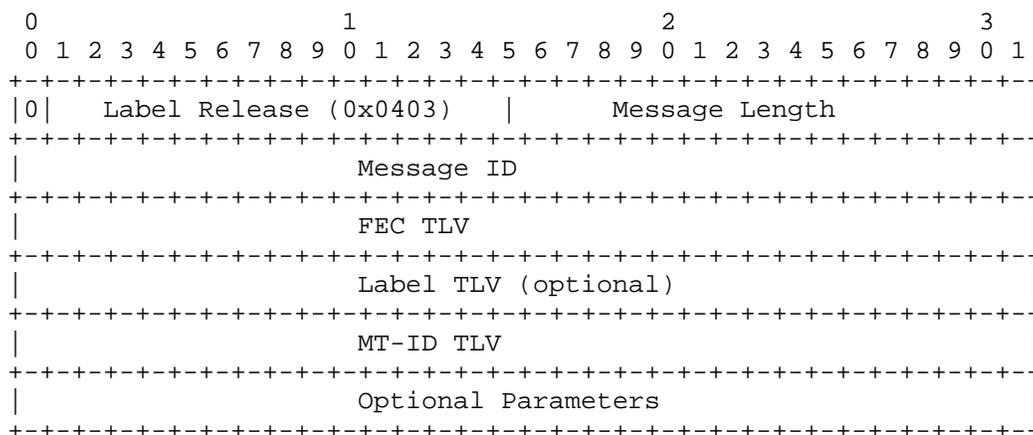


Figure 9: Label Release Message

9. Session Initialization Message with MT Capability

The session initialization message is extended to contain the LDP MT capability as an optional parameter. The extended session initialization message has the format below.

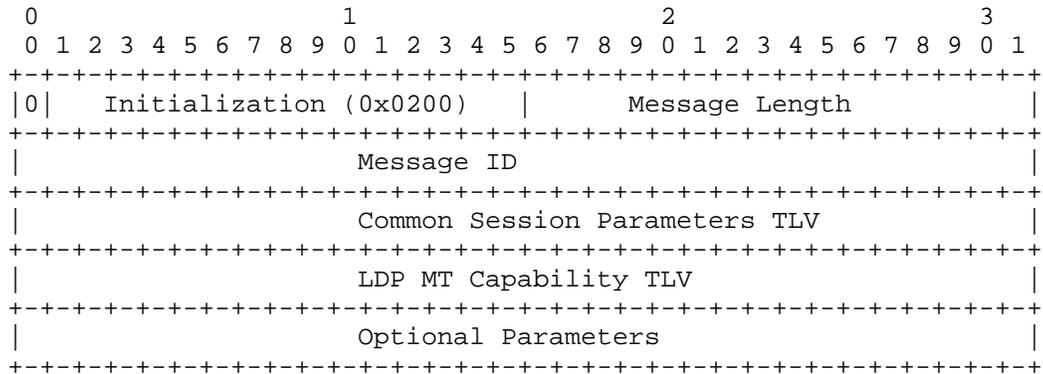


Figure 10: Session Initialization Message with MT Capability

10. MPLS Forwarding in MT

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

In MT based MPLS network, forwarding will be based not only on label, but also on MT-ID associated with the label. There are multiple options to do this. Below, we list the option preferred.

10.1. Use Label for (FEC, MT-ID) Tuple

We suggest is that MPLS forwarding for different topologies is implied by labels. This approach does not need any change to the existing MPLS hardware forwarding mechanism. It also resolves the forwarding issue that exists in IGP multi-topology forwarding when multiple topologies share an interface with overlay address space.

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 mechanism, for tuple (FEC-F, MT-ID-N1) and (FEC-F, MT-ID-N2), each LSR along the LSP path that is shared by topology MT-ID-N1 and MT-ID-N2 will allocate different labels to them. Thus two different Label Switching Paths will be created. One for (FEC-F, MT-ID-N1) and the other for (FEC-F, MT-ID-N2). The traffic for them will follow different Label Switching Paths (LSPs).

Note, in this mechanism, 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 accommodate MPLS MT. We will have different RIBs coresspoding to different MT IDs. But we will only need one LFIB.

Below is an example for MPLS forwarding:

```

RIB(x) for MT-IDx:
    FEC                NEXT HOP
    FECi(Destination A)  R1

RIB(y) for MT-IDy:
    FEC                NEXT HOP
    FECi(Destination A)  R2

LFIB:
    Ingress Label  Egress Label  NEXT HOP
    Lm             Lp             R1
    Ln             Lq             R2 (could be same as R1)

```

Figure 11: Forwarding Mechanism

11. 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.

12. IANA Considerations

TBD

13. Acknowledgement

The authors would like to thank Dan Tappan, Nabil Bitar, and Huang

Xin for their valuable comments on this draft.

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.
- [RFC3692] Narten, T., "Assigning Experimental and Testing Numbers Considered Useful", BCP 82, RFC 3692, January 2004.
- [RFC2434] Narten, T. and H. Alvestrand, "Guidelines for Writing an IANA Considerations Section in RFCs", BCP 26, RFC 2434, October 1998.
- [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-IS)", RFC 5120, February 2008.

14.2. 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

Emily Chen
Huawei Technology
No. 5 Street, Shangdi Information, Haidian
Beijing
China

Email: chenying220@huawei.com

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

Email: lilianyuan@chinamobile.com

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

Email: lilianyuan@chinamobile.com

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

Email: huanglu@chinamobile.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

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

Email: Ning.So@verizonbusiness.com

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

Email: pratiravi@juniper.com

