MPLS Working Group
Internet Draft
Intended status: Standards Track
Expires: April 2012

Dave Allan, Ed.
Ericsson

October 2011

Requirements and Framework for Unified MPLS Sub-Network Interconnection

draft-allan-unified-mpls-ic-req-frmwk-00

Abstract

The definition of a transport profile for MPLS means that MPLS network architectures will emerge that combines both managed and control plane driven MPLS sub-networks and requires interconnection of same to achieve a unified MPLS architecture.

This document generalizes the problem of sub-network interconnect, discusses issues in general and suggests ways forward.

Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC2119 [1].

Status of this Memo

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

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

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

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

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

This Internet-Draft will expire in January 2012.

Table of Contents

1. Introduction

Networks that provide an end-to-end service infrastructure are typically deployed as multiple domains or sub-networks in order to scale. These different domains may be operated by different technologies using different control plane technology (e.g. management driven control plane (centralized) or distributed control plane).

Within the MPLS control plane there exist a number of functional behaviors that are typically associated with a single control protocol and function autonomously from the other control protocols. That is to say when a labeled layer and associated control protocol is overlaid on another, there is frequently no operational coupling between them. When two control protocols are operated side by side the same is true (e.g. ships in the night). An example of the former would be pseudo wires over a PSN. An example of the latter would be L2 and L3 virtual private networks (VPNs) sharing a common packet switched network (PSN).

The introduction of transport functions to MPLS and the intent to deploy, merge or otherwise combine networks that will concatenate sub-networks that have different operational characteristics and/or control planes (including management) introduces the requirement to integrate operations across multiple sub-networks. This frequently is manifested in requirements on nodes at sub-network boundaries and/or alignment of identifiers in order to construct a unified end-to-end MPLS based service plane.

By having a unified MPLS based service plane, a network operator is able to provide services across different MPLS domains instrumented with common management and control functionality in order to provide scalable service offerings on a common MPLS technology base.

This document is a framework that postulates models and functional requirements for sub-network interconnect to achieve a unified end to end MPLS based service plane or "Unified MPLS".

2. Conventions used in this document

2.1. Terminology

Binding: Is the mechanism for association and interconnection of label switched path (LSP) or pseudo wire (PW) segments that exist in different sub-networks.

Section: As per RFC 5960[17].

Segment: A part of a PW or LSP that has one or more contiguous sections in a common sub-network. This usage is as per RFC 6372[18].

Sub-network: A portion of a larger MPLS network that is operated by a single system and whose boundaries are set by the scope of the system. The system may be a control plane or management system. Similarly it could be a sub-layer stacked on another sub-network.

2.2. Acronyms

LIB - label information base

LSP - label switched path

MEG - Maintenance Entity Group

MEP - Maintenance Entity Group End Point

MIP - Maintenance Entity Group Intermediate Point

PSN - packet switched network

PW - pseudo wire

SPME - Sub-path Maintenance Entity

3. Sub-Network Interconnect Scenarios

The combination of MPLS label swapping and label stacking makes it possible to consider a number of atomic interconnect scenarios, briefly summarized here:

Peer Model - is the scenario when an LSP or PW has segments in different sub-networks.

Segmentation Model - is the scenario when a client LSP or PW with common establishment and maintenance procedures has sections in separate sub-networks. This model can recurse arbitrarily.

Overlay Model - is the simplest case of the segmentation model in which a section of an LSP or PW is a distinct sub-network.

Termination Model - is the scenario where a non-MPLS or PW transfer function separates the sub-networks. The termination model logically isolates the sub-networks and is included simply for completeness.

4. Sub-Network Interconnect Mechanisms

There are two interconnect mechanisms considered. The first (Border node) postulates a node that spans two sub-networks and both likely operated by a single provider. The second (Border link) postulates a link connecting sub-networks of two distinct organizations within a provider or two distinct providers.

4.1. Border Node

A border node is one that implements, interconnects and interworks LSPs or PWs with segments or sections in different sub-networks. The interworking function at the border node will either terminate, swap or encapsulate labels.

4.2. Border Link

A border link is the case whereby two border nodes are connected back to back in a sub-network that consists of a single link.

5. Sub-network types

In the current MPLS architecture the following sub-network types exist:

5.1. Infrastructure sub-network types

MPLS-TD: Topology driven MPLS. LSP setup is via the LDP protocol [6] with path routing determined by the IGP. ECMP, merging and PHP are are included in the set of possible dataplane behaviors. P2mp LSPs are supported by mLDP [7].

MPLS-TE: LSP setup is via the RSVP-TE protocol[8] with path routing determined by a TE enabled IGP (e.g. OSPF-TE) according to bandwidth and QoS constraints. PHP is included in the set of dataplane behaviors. Bandwidth allocation, class of service or priotity(PHB) are typically part of traffic handling.

Managed MPLS-TP: LSP setup is via management action. LSPs are purely connection oriented p2p or p2mp.

Signalled MPLS-TP: LSP setup is via RSVP-TE GMPLS[9]. LSPs are purely connection oriented p2p or p2mp.

5.2. Service Sub-network types

L3VPN: VPN setup is via the BGP protocol as per RFC 4364[10]. Note that this can be an IP-VPN (via IGP peering with the VRF) or an MPLS-VPN (via a combination of IGP and MPLS CP peering with the VRF).

L2VPN: VPN setup is via the PW Control Protocol per RFC 4447 (LDP in targeted mode) or BGP protocols. A broadcast domain is emulated which will have the effect of limiting sub-network interconnect to a single point to avoid loops.

VPWS: VPWS setup is via the PW Control Protocol per rfc 4447 (LDP in targeted mode).

6. Issues & Requirements

In theory, any ingress label can be mapped to one or more egress labels or label stack permutations via the ILM to NHLFE mapping defined in RFC 3031[2]. Further one carrier"s infrastructure layer may be a client of another carrier"s infrastructure. More considerations need to come into play in order to produce a tractable set of sub-network interworking scenarios. The following is a partial list of some of the issues to be considered and/or addressed:

6.1. Alignment of OAM functionality

Not all OAM encapsulations guarantee fate sharing with the LSP of interest across all of the sub-network types in MPLS. This not only means that failures may not be detected or detected in a timely manner, it also means that "false positives" are a possibility as failures may occur on the path taken by the OAM PDUs.

Any OAM encapsulation using a reserved label, e.g. the GAL[12], or Router Alert as used by VCCV type 2[13], or without a PW control word will not have predictable control over fate sharing with normal payload for any LSP or PW that has a section that transits a MPLS-TD subnetwork that implements ECMP[11].

A separate issue is interconnecting subnetworks where the LSPs have a different cardinality of end points (e.g. concatenating mp2p to p2p), implying a different number of maintenance entities than would be suggested by an implementation dimensioned to a single subnetwork"s characteristics.

6.2. OAM identifier mismatches

MEG, MEP, MIP and nodal addressing will not pose identifier mismatch problems. Where such problems will arise is in the use of RFC 4379 LSP Ping [3]. This is because LSP-PING uses identifiers associated with a specific sub-network type in the FEC stack as part of the processing to detect inconsistencies between the control plane and the data plane.

[Issue: The on demand cv draft provides for the Static LSP and Static PW TLVs for the FEC stack which allows intermediate nodes to validate the FEC stack against the MEG ID for the local MIP. The applicability for this should be generalized such that it can be used end to end across domains. This will also raise the problem of disseminating MEG information to non-transport subnetworks as not all defined MPLS sub-network types use the current fields in the IP based LSP MEG ID]

6.3. OAM encapsulation

The MPLS architecture permits multiple OAM encapsulations that may or may nor have an IP header. Any interconnect mechanism needs to be able to align not only capability but encapsulations end to end.

6.4. Protection Mechanisms

An MPLS LSP or PW sub-network may be made resilient by any number of mechanisms. There is also three scenarios of note, end to end protection, end to end restoration and sub-network protection.

End to end protection offers minimal complications in sub-network interconnect as the interworking functions is common to that of the unprotected case, that is to say transit nodes do not participate in protection switching.

Sub-network protection is universally offered by the use of mechanisms that operate within the level such as detours [19] and may require label merging at the border node. Mechanisms that operate at nested MPLS label levels (e.g. SPMEs or FRR facility protection) have no impact on sub-network interconnect.

End to end restoration is a bit more complicated as it involves coordinating dynamic action between sub-networks.

It also becomes possible to consider sub-network restoration with many of the same considerations as path maintenance and re-sizing.

6.5. Label space management

The MPLS architecture has always been based around local administration of a node"s label space. As such mechanisms to partition the label space between multiple administrative entities is not currently supported and would be difficult to retrofit.

A consequence of this is that a border node is potentially required to provide labels from a common pool to both a control and management plane, e.g. a management system be required to obtain label values from the node prior to populating the LIB vs. being delegated a pool. This suggests that such a mechanism be carried forward for all managed nodes such that only a single mechanism need be implemented. However this is an implementation decision.

6.6. Path maintenance and re-sizing

It must be possible to make operational modifications to a path segment in a hitless fashion. The normal procedure for MPLS-TE is known as "make before break". This gives rise to two scenarios, the first is end to end "make before break", and the other is make before break confined to a sub-network with the border node as a pinned waypoint. This means the design of the inter-sub-network binding information permit make before break modification of one segment of the LSP.

6.7. Sub-network migration

The practical considerations are documented in RFC 5950[4] and by reference RFC 5493[5].

6.8. (Non)Interworking of DP and CP notifications

Within the MPLS architecture there are techniques for propagating the status of adjacent sections of either a native service or PW section in both the data plane and the control plane. One example (documenting both) is RFC 6310[14].

When concatenating subnetworks the interworking of dataplane fault notifications [15] or protection switching coordination [16] and control plane indications will not be possible. The reason is that data plane indications flow end to end on a labeled path therefore will not be visible to border nodes, a requirement to enable interworking of dataplane notifications with the control plane in any useful form.

When connecting a subnetwork restricted to data plane only notifications to a subnetwork that will support either dataplane or

control plane notifications, the border node will be required to negotiate exclusive use of dataplane notifications in any control plane signaling during the path setup. This will have implications in both the interconnect data model, and potential enhancements to signaling.

7. Operational Decoupling

The objective of any sub-network interconnect solution is to decouple the operation of the interconnected systems in order to minimize any dependencies.

The sub-network interconnect must accommodate interconnecting LSPs and PWs with different establishment and persistency characteristics. This is determined by whether the LSP, PW or segment is provisioned or signaled, where from a persistency point of view, a provisioned entry is permanent and exists until removed by management action, while a signaled entity fate shares with a control plane adjacency and may come and go during the life time of the inter sub-network binding.

The state present at a border node to bind the LSP or PW spanning the subnetworks together should exist independently of the characteristics of the LSPs or associated control or management planes.

This also requires a level of indirection such that the management action is decoupled from the mechanics of label assignment in each sub-network and may work with sub-network resiliency mechanisms.

So the state "connect whatever label from sub-network A associated with FOO to whatever label from sub-network B is associated with BAR" should be persistent.

8. Acknowledgments

Loa Andersson, Eric Gray, David Sinicrope and Greg Mirsky contributed to the development of this document.

9. IANA Considerations

This document does not require IANA action.

10. Security Considerations

For a future version of this document.

11. References

11.1. Informative References

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

[2] Rosen et.al. Multiprotocol Label Switching Architecture, IETF RFC 3031, January 2001

[3] Kompella et.al. Detecting Multi-Protocol Label Switched (MPLS) Data Plane Failures, IETF RFC 4379, February 2006

[4] Mansfield et. al. Network Management Framework for MPLS-based Transport Networks, IETF RFC 5950, September 2010

[5] Caviglia, D., Bramanti, D., Li, D., and D. McDysan, "Requirements for the Conversion between Permanent Connections and Switched Connections in a Generalized Multiprotocol Label Switching (GMPLS) Network", RFC 5493, April 2009.

[6] Andersson et.al. LDP Specification, IETF RFC 5036, October 2007

[7] Minei, I. et.al. Label Distribution Protocol Extensions for Point-to-Multipoint and Multipoint-to-Multipoint Label Switched Paths, IETF work in progress, draft-ietf-mpls-ldp-p2mp-15

[8] Awduche et.al. RSVP-TE: Extensions to RSVP for LSP Tunnels, IETF RFC 3209, December 2001

[9] Berger et.al. Generalized Multi-Protocol Label Switching (GMPLS) Signaling Functional Description, IETF RFC 3471, January 2003

[10] Rosen et.al. BGP/MPLS IP Virtual Private Networks (VPNs), IETF RFC 4364, February 2006

[11] Swallow et.al. Avoiding Equal Cost Multipath Treatment in MPLS Networks, IETF RFC 4928, June 2007

[12] Bocci et.al. MPLS Generic Associated Channel, IETF RFC 5586, June 2009

[13] Nadeau et.al Pseudowire Virtual Circuit Connectivity Verification (VCCV) A Control Channel for Pseudowires, IETF RFC 5085, December 2007

[14] Aissaoui et.al. Pseudowire (PW) Operations, Administration, and Maintenance (OAM) Message Mapping, IETF RFC 6310, July 2011

[15] Swallow et.al. MPLS Fault Management OAM, IETF work in progress, draft-ietf-mpls-tp-fault-07, September 2011

[16] Bryant et.al. MPLS-TP Linear Protection, IETF work in progress, draft-ietf-mpls-tp-linear-protection-09.txt, August 2011

[17] Frost et.al. MPLS Transport Profile Data Plane Architecture, IETF RFC 5960, August 2010

[18] Sprecher et.al. MPLS Transport Profile (MPLS-TP) Survivability Framework, IETF RFC 6372, September 2011

[19] Pan et.al. Fast Reroute Extensions to RSVP-TE for LSP Tunnels, IETF RFC 4090, May 2005

Authors' Addresses

Dave Allan
Ericsson
Email: david.i.allan@ericsson.com

Network Working Group
Internet-Draft
Intended status: Informational
Expires: May 1, 2012

T. Beckhaus
Deutsche Telekom AG
B. Decraene
France Telecom
K. Tiruveedhula
M. Konstantynowicz
Juniper Networks
L. Martini
Cisco Systems, Inc.
October 29, 2011

LDP Downstream-on-Demand in Seamless MPLS
draft-beckhaus-ldp-dod-01

Abstract

Seamless MPLS design enables a single IP/MPLS network to scale over core, metro and access parts of a large packet network infrastructure using standardized IP/MPLS protocols. One of the key goals of Seamless MPLS is to meet requirements specific to access, including high number of devices, their position in network topology and their compute and memory constraints that limit the amount of state access devices can hold.This can be achieved with LDP Downstream-on-Demand (LDP DoD) label advertisement. This document describes LDP DoD use cases and lists required LDP DoD procedures in the context of Seamless MPLS design.

In addition, a new optional TLV type in the LDP label request message is defined for fast-up convergence.

Requirements Language

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

Status of this Memo

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

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

Internet-Drafts are draft documents valid for a maximum of six months

and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on May 1, 2012.

Copyright Notice

Table of Contents

1. Introduction

Seamless MPLS design [I-D.ietf-mpls-seamless-mpls] enables a single IP/MPLS network to scale over core, metro and access parts of a large packet network infrastructure using standardized IP/MPLS protocols. One of the key goals of Seamless MPLS is to meet requirements specific to access, including high number of devices, their position in network topology and their compute and memory constraints that limit the amount of state access devices can hold.

In general MPLS routers implement either LDP or RSVP for MPLS label distribution. The focus of this document is on LDP, as Seamless MPLS design does not include a requirement for general purpose explicit traffic engineering and bandwidth reservation. This document is focusing on the unicast connectivity only. Multicast connectivity is subject for further study.

In Seamless MPLS design [I-D.ietf-mpls-seamless-mpls], IP/MPLS protocol optimization is possible due to a relatively simple access network topologies. Examples of such topologies involving access nodes (AN) and aggregation nodes (AGN) include:

a. A single AN homed to a single AGN.

b. A single AN dual-homed to two AGNs.

c. Multiple ANs daisy-chained via a hub-AN to a single AGN.

d. Multiple ANs daisy-chained via a hub-AN to two AGNs.

e. Two ANs dual-homed to two AGNs.

f. Multiple ANs chained in a ring and dual-homed to two AGNs.

The amount of IP RIB and FIB state on ANs can be easily controlled in the listed access topologies by using simple IP routing configuration with either static routes or dedicated access IGP. Note that in all of the above topologies AGNs act as the access border routers (access ABRs) connecting the access topology to the rest of the network. Hence in many cases it is sufficient for ANs to have a default route pointing towards AGNs in order to achieve complete network connectivity from ANs to the network.

The amount of MPLS forwarding state however requires additional consideration. In general MPLS routers implement LDP Downstream Unsolicited (LDP DU) label advertisement [RFC5036] and advertise MPLS labels for all valid routes in their RIB. This is seen as a very insufficient approach for ANs, as they only require a small subset of

the total routes (and associated labels) based on the required connectivity for the provisioned services. And although filters can be applied to those LDP DU labels advertisements, it is not seen as a suitable tool to facilitate any-to-any AN-driven connectivity between access and the rest of the MPLS network.

This document describes an access node driven "subscription model" for label distribution in the access. The approach relies on the standard LDP Downstream-on-Demand (LDP DoD) label advertisements as specified in [RFC5036]. LDP DoD enables on-demand 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 LDP specification [RFC5036]. This is due to the fact that the originally LDP DoD advertisement mode was aimed mainly at ATM and Frame Relay MPLS implementations, where conserving label space used on the links was essential for compatibility with ATM and Frame Relay LSRs.

The following sections describe a set of reference access topologies considered for LDP DoD usage and their associated IP routing configurations, followed by LDP DoD use cases and LDP DoD procedures in the context of Seamless MPLS design.

2. Reference Topologies

LDP DoD use cases are described in the context of a generic reference end-to-end network topology based on Seamless MPLS design [I-D.ietf-mpls-seamless-mpls] shown in Figure 1

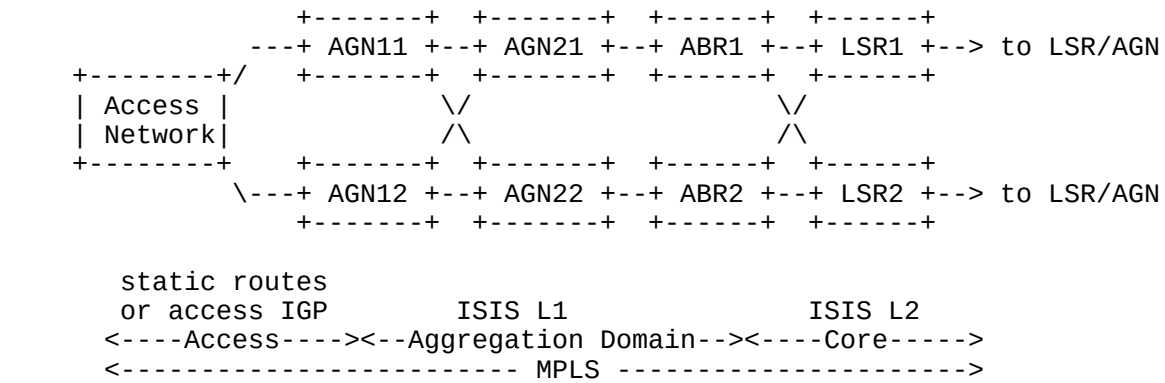


Figure 1: Seamless MPLS end-to-end reference network topology.

The access network is either single or dual homed to AGN1x, with either a single or multiple parallel links to AGN1x.

Seamless MPLS access network topologies can range from a single- or dual-homed access node to a chain or ring of access nodes, and use either static routing or access IGP. The following sections describe reference access topologies in more detail.

2.1. Access Topologies with Static Routing

In most cases access nodes connect to the rest of the network using very simple topologies. Here static routing is sufficient to provide the required IP connectivity. The following topologies are considered for use with static routing and LDP DoD:

a. [I1] topology - a single AN homed to a single AGN.

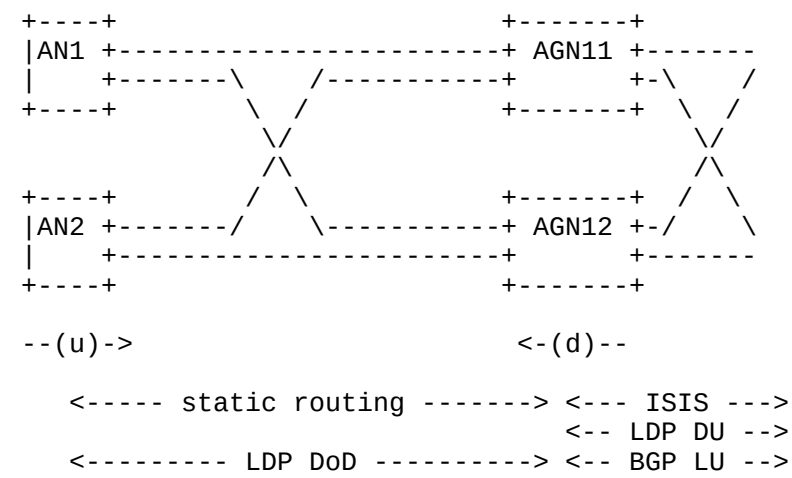
b. [I] topology - multiple ANs daisy-chained to a single AGN.

c. [V] topology - a single AN dual-homed to two AGNs.

d. [U2] topology - two ANs dual-homed to two AGNs.

e. [Y] topology - multiple ANs daisy-chained to two AGNs.

The reference static routing and LDP configuration for [V] access topology is shown in Figure 2. The same static routing and LDP configuration also applies to [I1] topology.

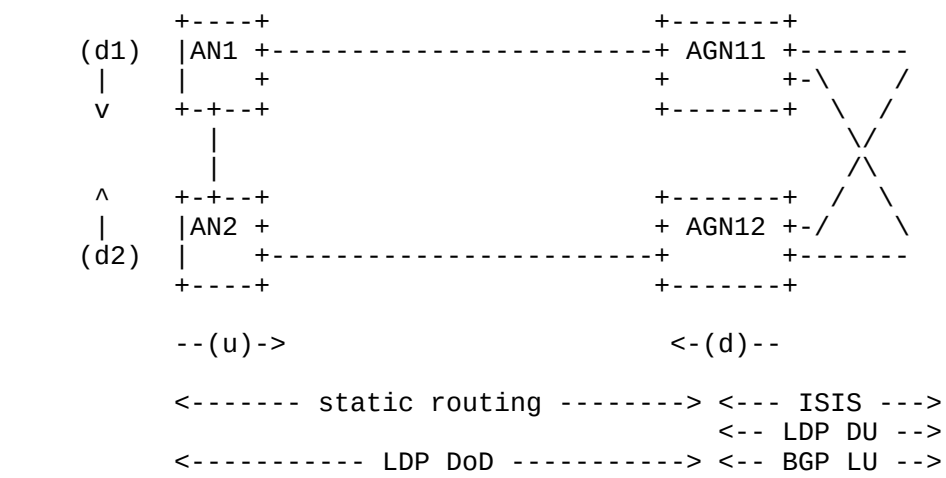


(u) static routes: 0/0 default, (optional) /32 or /128 destinations
(d) static routes: /32 or /128 AN loopbacks

Figure 2: [V] access topology with static routes.

In line with the Seamless MPLS design, static routes configured on AGN1x and pointing towards the access network are redistributed in either ISIS or BGP labeled unicast (BGP-LU) [RFC3107].

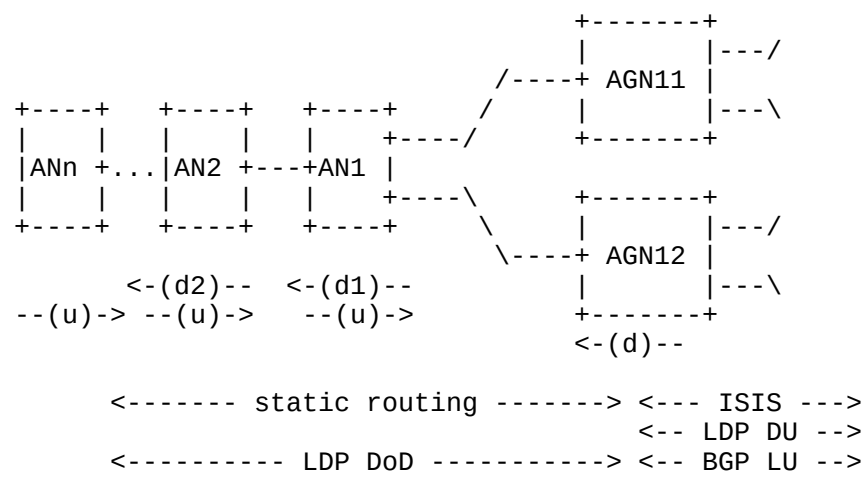
The reference static routing and LDP configuration for [U2] access topology is shown in Figure 3.



(u) static route 0/0 default (/32 or /128 destinations optional)
(d) static route for /32 or /128 AN loopbacks
(d1) static route for /32 or /128 AN2 loopback and 0/0 default with lower preference
(d2) static route for /32 or /128 AN1 loopback and 0/0 default with lower preference

Figure 3: [U2] access topology with static routes.

The reference static routing and LDP configuration for [Y] access topology is shown in Figure 4. The same static routing and LDP configuration also applies to [I] topology.



(u) static routes: 0/0 default, (optional) /32 or /128 destinations
(d) static routes: /32 or /128 AN loopbacks [1..n]
(d1) static routes: /32 or /128 AN loopbacks [2..n]
(d2) static routes: /32 or /128 AN loopbacks [3..n]

Figure 4: [Y] access topology with static routes.

Note that in all of the above topologies parallel ECMP (or L2 LAG) links can be used between the nodes.

ANs support Inter-area LDP [RFC5283] in order to use the IP default route to match the LDP FEC advertised by AGN1x and other ANs.

2.2. Access Topologies with Access IGP

A dedicated access IGP instance is used in the access network to perform the internal routing between AGN1x and connected AN devices. Example of such IGP could be ISIS, OSPFv2&v3, RIPv2&RIPng. This access IGP instance is distinct from the IGP of the aggegation domain.

The following topologies are considered for use with access IGP routing and LDP DoD:

a. [U] topology - multiple ANs chained in an open ring and dual-homed to two AGNs.

b. [Y] topology - multiple ANs daisy-chained via a hub-AN to two AGNs.

The reference access IGP and LDP configuration for [U] access

topology is shown in Figure 5.

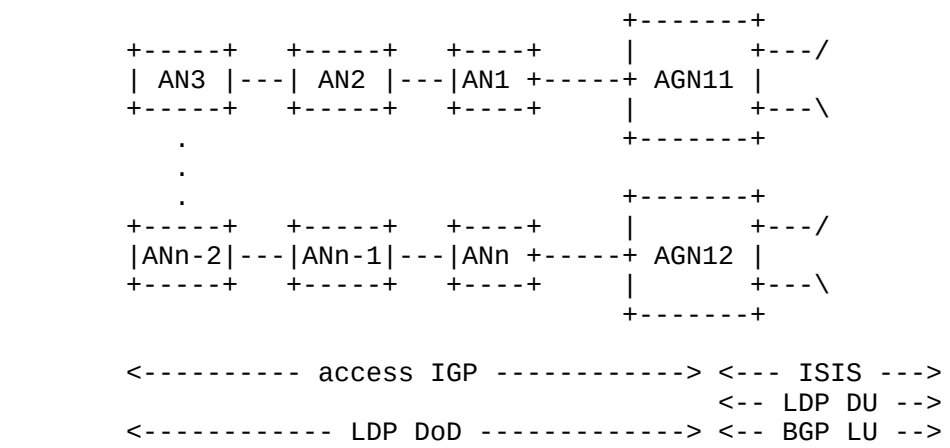


Figure 5: [U] access topology with access IGP.

The reference access IGP and LDP configuration for [Y] access topology is shown in Figure 6.

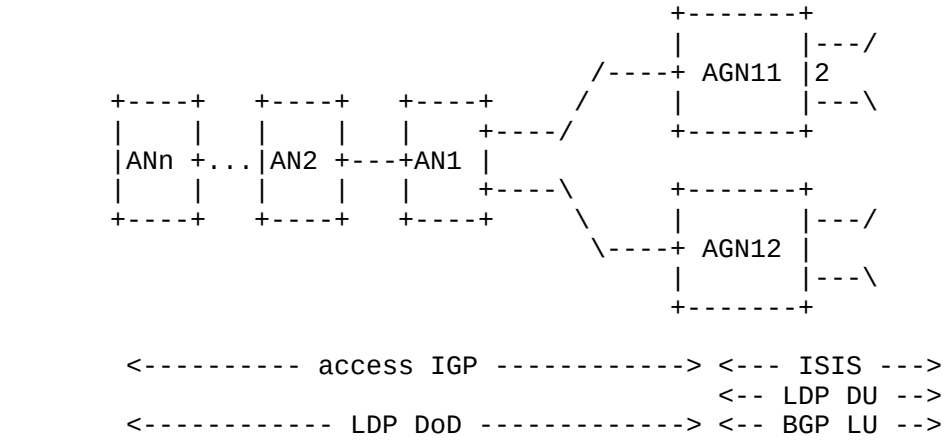


Figure 6: [Y] access topology with access IGP.

Note that in all of the above topologies parallel ECMP (or L2 LAG) links can be used between the nodes.

In both of the above topologies, ANs (ANn ... AN1) and AGN1x share the access IGP and advertise their IPv4 and IPv6 loopbacks and link addresses. AGN1x advertise a default route into the access IGP.

ANs support Inter-area LDP [RFC5283] in order to use the IP default route for matching the LDP FECs advertised by AGN1x or other ANs.

3. LDP DoD Use Cases

LDP DoD operation is driven by Seamless MPLS use cases. This section illustrates these use cases focusing on services provisioned on the access nodes and clarifies expected LDP DoD operation on the AN and AGN1x devices. Two representative service types are used to illustrate the service use cases: MPLS PWE3 [RFC4447] and BGP/MPLS IPVPN [RFC4364].

Described LDP DoD operations apply equally to all reference access topologies described in Section 2. Operations that are specific to certain access topologies are called out explicitly.

References to upstream and downstream nodes are made in line with the definition of upstream and downstream LSR [RFC3031].

This document is focusing on IPv4 LDP DoD procedures. Similar procedures are required for IPv6 LDP DoD, however some extension specific to IPv6 are likely to 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. Initial Network Setup

An access node is commissioned without any services provisioned on it. The AN may request labels for loopback addresses of any AN, AGN or other nodes within Seamless MPLS network for operational and management purposes. It is assumed that AGN1x has required IP/MPLS configuration for network-side connectivity in line with Seamless MPLS design [I-D.ietf-mpls-seamless-mpls].

LDP sessions are configured between adjacent ANs and AGN1x using their respective loopback addresses.

3.1.1. AN with Static Routing

If access static routing is used, ANs are provisioned with the following static IP routing entries (topology references from Section 2 are listed in square brackets):

a. [I1, V, U2] - Static default route 0/0 pointing to links connected to AGN1x. Requires support for Inter-area LDP [RFC5283].

b. [U2] - Static /32 or /128 routes pointing to the other AN. Lower preference static default route 0/0 pointing to links connected to the other AN. Requires support for Inter-area LDP [RFC5283].

c. [I, Y] - Static default route 0/0 pointing to links leading towards AGN1x. Requires support for Inter-area LDP [RFC5283].

d. [I, Y] - Static /32 or /128 routes to all ANs in the daisy-chain pointing to links towards those ANs.

e. [I1, V, U2] - Optional - Static /32 or /128 routes for specific nodes within Seamless MPLS network, pointing to links connected to AGN1x.

f. [I, Y] - Optional - Static /32 or /128 routes for specific nodes within the Seamless MPLS network, pointing to links leading towards AGN1x.

Upstream AN/AGN1x should request labels over LDP DoD session(s) from downstream AN/AGN1x for configured static routes if those static routes are configured with LDP DoD request policy and if they are pointing to a next-hop selected by routing. It is expected that all configured /32 and /128 static routes to be used for LDP DoD are configured with such policy on AN/AGN1x.

Downstream AN/AGN1x should respond to the label request from the upstream AN/AGN1x with a label mapping (if requested route is present in its RIB, and there is a valid label binding from its downstream), and must install the advertised label as an incoming label in its label table (LIB) and its forwarding table (LFIB). Upstream AN/AGN1x must also install the received label as an outgoing label in their LIB and LFIB. If the downstream AN/AGN1x does have the route present in its RIB, but does not have a valid label binding from its downstream, it should forward the request to its downstream.

In order to facilitate ECMP and IPFRR LFA local-repair, the upstream AN/AGN1x must also send LDP DoD label requests to alternate next-hops per its RIB, and install received labels as alternate entries in its LIB and LFIB.

AGN1x node on the network side may use BGP labeled unicast [RFC3107] in line with the Seamless MPLS design [I-D.ietf-mpls-seamless-mpls]. In such a case AGN1x will be redistributing its static routes pointing to local ANs into BGP labeled unicast to facilitate network-to-access traffic flows. Likewise, to facilitate access-to-network traffic flows, AGN1x will be responding to access-originated LDP DoD label requests with label mappings based on its BGP labeled unicast reachability for requested FECs.

3.1.2. AN with Access IGP

If access IGP is used, AN(s) advertise their loopbacks over the access IGP with configured metrics. AGN1x advertise a default route over the access IGP.

Similarly to the static route case, upstream AN/AGN1x should request labels over LDP DoD session(s) from downstream AN/AGN1x for all /32 or /128 routes received over the access IGP.

Identically to the static route case, downstream AN/AGN1x should respond to the label request from the upstream AN/AGN1x with a label mapping (if the requested route is present in its RIB, and there is a valid label binding from its downstream), and must install the advertised label as an incoming label in its LIB and LFIB. Upstream AN/AGN1x must also install the received label as an outgoing label in their LIB and LFIB.

Identically to the static route case, in order to facilitate ECMP and IPFRR LFA local-repair, upstream AN/AGN1x must also send LDP DoD label requests to alternate next-hops per its RIB, and install received labels as alternate entries in its LIB and LFIB.

AGN1x node on the network side may use BGP labeled unicast [RFC3107] in line with Seamless MPLS design [I-D.ietf-mpls-seamless-mpls]. In such case AGN1x will be redistributing routes received over the access IGP (and pointing to local ANs), into BGP labeled unicast to facilitate network-to-access traffic flows. Likewise, to facilitate access-to-network traffic flows AGN1x will be responding to access originated LDP DoD label requests with label mappings based on its BGP labeled unicast reachability for requested FECs.

3.2. Service Provisioning and Activation

Following the initial setup phase described in Section 3.1, a specific access node, referred to as AN*, is provisioned with a network service. AN* relies on LDP DoD to request the required MPLS LSP(s) label(s) from downstream AN/AGN1x node(s). Note that LDP DoD operations are service agnostic, that is, they are the same independently of the services provisioned on the AN*.

For illustration purposes two service types are described: MPLS PWE3 [RFC4447] service and BGP/MPLS IPVPN [RFC4364].

MPLS PWE3 service - for description simplicity it is assumed that a single segment pseudowire is signaled using targeted LDP FEC128 (0x80), and it is provisioned with the pseudowire ID and the loopback IPv4 address of the destination node. The following IP/MPLS

operations need to be completed on the AN* to successfully establish such PWE3 service:

a. LSP labels for destination /32 FEC (outgoing label) and the local /32 loopback (incoming label) need to be signaled using LDP DoD.

b. Targeted LDP session over an associated TCP/IP connection needs to be established to the PWE3 destination PE. This is triggered by either an explicit targeted LDP session configuration on the AN* or automatically at the time of provisioning the PWE3 instance.

c. Local and remote PWE3 labels for specific FEC128 PW ID need to be signaled using targeted LDP and PWE3 signaling procedures [RFC4447].

d. Upon successful completion of the above operations, AN* programs its RIB/LIB and LFIB tables, and activates the MPLS PWE3 service.

Note - only minimum operations applicable to service connectivity have been listed. Other non IP/MPLS connectivity operations that may be required for successful service provisioning and activation are out of scope in this document.

BGP/MPLS IPVPN service - for description simplicity it is assumed that AN* is provisioned with a unicast IPv4 IPVPN service (VPNv4 for short) [RFC4364]. The following IP/MPLS operations need to be completed on the AN* to successfully establish VPNv4 service:

a. BGP peering sessions with associated TCP/IP connections need to be established with the remote destination VPNv4 PEs or Route Reflectors.

b. Based on configured BGP policies, VPNv4 BGP NLRIs need to be exchanged between AN* and its BGP peers.

c. Based on configured BGP policies, VPNv4 routes need to be installed in the AN* VRF RIB and FIB, with corresponding BGP next-hops.

d. LSP labels for destination BGP next-hop /32 FEC (outgoing label) and the local /32 loopback (incoming label) need to be signaled using LDP DoD.

e. Upon successful completion of above operations, AN* programs its RIB/LIB and LFIB tables, and activates the BGP/MPLS IPVPN service.

Note - only minimum operations applicable to service connectivity have been listed. Other non IP/MPLS connectivity operations that may be required for successful service provisioning are out of scope in this document.

To establish an LSP for destination /32 FEC for any of the above services, AN* looks up its local routing table for a matching route, selects the best next-hop(s) and associated outgoing link(s).

If a label for this /32 FEC is not already installed based on the configured static route with LDP DoD request policy or access IGP RIB entry, AN* must send an LDP DoD label mapping request. Downstream AN/AGN1x LSR(s) checks its RIB for presence of the requested /32 and associated valid outgoing label binding, and if both are present, replies with its label for this FEC and installs this label as incoming in its LIB and LFIB. Upon receiving the label mapping the AN* must accept this label based on the exact route match of advertised FEC and route entry in its RIB or based on the longest match in line with Inter-area LDP [RFC5283]. If the AN* accepts the label it must install it as an outgoing label in its LIB and LFIB.

In access topologies [V] and [Y], if AN* is dual homed to two AGN1x and routing entries for these AGN1x are configured as equal cost paths, AN* must send LDP DoD label requests to both AGN1x devices and install all received labels in its LIB and LFIB.

In order for AN* to implement IPFRR LFA local-repair, AN* must also send LDP DoD label requests to alternate next-hops per its RIB, and install received labels as alternate entries in its LIB and LFIB.

When forwarding PWE3 or VPNv4 packets AN* chooses the LSP label based on the locally configured static /32 or default route, or default route signaled via access IGP. If a route is reachable via multiple interfaces to AGN1x nodes and the route has multiple equal cost paths, AN* must implement Equal Cost Multi-Path (ECMP) functionality. This involves AN* using hash-based load-balancing mechanism and sending the PWE3 or VPNv4 packets in a flow-aware manner with appropriate LSP labels via all equal cost links.

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

AGN1x node on the network side may use BGP labeled unicast [RFC3107] in line with Seamless MPLS design [I-D.ietf-mpls-seamless-mpls]. In such case AGN1x will be redistributing its static routes (or routes received from the access IGP) pointing to local ANs into BGP labeled

unicast to facilitate network-to-access traffic flows. Likewise, to facilitate access-to-network traffic flows AGN1x will be responding to access originated LDP DoD label requests with label mappings based on its BGP labeled unicast reachability for requested FECs.

3.3. Service Changes and Decommissioning

Whenever AN* service gets decommissioned or changed and connectivity to specific destination is not longer required, the associated MPLS LSP label resources should be released on AN*.

MPLS PWE3 service - if the PWE3 service gets decommissioned and it is the last PWE3 to a specific destination node, the targeted LDP session is not longer needed and should be terminated (automatically or by configuration). The MPLS LSP(s) to that destination is no longer needed either.

BGP/MPLS IPVPN service - deletion of a specific VPNv4 (VRF) instance, local or remote re-configuration may result in specific BGP next-hop(s) being no longer needed. The MPLS LSP(s) to that destination is no longer needed either.

In all of the above cases the following LDP DoD related operations apply:

- If the /32 FEC label for the aforementioned destination node was originally requested based on either tLDP session configuration and default route or required BGP next-hop and default route, AN* should delete the label from its LIB and LFIB, and release it from downstream AN/AGN1x by using LDP DoD procedures.

- If the /32 FEC label was originally requested based on the static /32 route configuration with LDP DoD request policy, the label must be retained by AN*.

3.4. Service Failure

A service instance may stop being operational due to a local or remote service failure event.

In general, unless the service failure event modifies required MPLS connectivity, there should be no impact on the LDP DoD operation.

If the service failure event does modify the required MPLS connectivity, LDP DoD operations apply as described in Section 3.2 and Section 3.3.

3.5. Network Transport Failure

A number of different network events can impact services on AN*. The following sections describe network event types that impact LDP DoD operation on AN and AGN1x nodes.

3.5.1. General Notes

If service on any of the ANs is affected by any network failure and there is no network redundancy, the service must go into a failure state. When the network failure is recovered from, the service must be re-established automatically.

The following additional LDP-related functions should be supported to comply with Seamless MPLS [I-D.ietf-mpls-seamless-mpls] fast service restoration requirements as follows:

a. Local-repair - AN and AGN1x should support local-repair for adjacent link or node failure for access-to-network, network-to-access and access-to-access traffic flows. Local-repair should be implemented by using either IPFRR LDP LFA, simple ECMP or primary/backup switchover upon failure detection.

b. LDP session protection - LDP sessions should be configured with LDP session protection to avoid delay upon the recovery from link failure. LDP session protection ensures that FEC label binding is maintained in the control plane as long as LDP session stays up.

c. IGP-LDP synchronization - If access IGP is used, LDP sessions between ANs, and between ANs and AGN1x, should be configured with IGP-LDP synchronization to avoid unnecessary traffic loss in case the access IGP converged before LDP and there is no LDP label binding to the downstream best next-hop.

3.5.2. AN Node Failure

AN node fails and all links to adjacent nodes go down.

Adjacent AN/AGN1x nodes remove all routes pointing to the failed link(s) from their RIB tables (including /32 loopback belonging to the failed AN and any other routes reachable via the failed AN). This in turn triggers the removal of associated outgoing /32 FEC labels from their LIB and LFIB tables.

If access IGP is used, the AN node failure will be propagated via IGP link updates across the access topology.

If a specific /32 FEC(s) is not reachable anymore from those AN/ AGN1x, they must also send LDP label withdraw to their upstream LSRs to notify about the failure, and remove the associated incoming label(s) from their LIB and LFIB tables. Upstream LSRs upon receiving label withdraw should remove the signaled labels from their LIB/LFIB tables, and propagate LDP label withdraw across their upstream LDP DoD sessions.

In [U] topology there may be an alternative path to routes previously reachable via the failed AN node. In this case adjacent AN/AGN1x should invoke local-repair (IPFRR LFA, ECMP) and switchover to alternate next-hop to reach those routes.

AGN1x gets notified about the AN failure via either access IGP (if used) and/or cascaded LDP DoD label withdraw(s). AGN1x must implement all relevant global-repair IP/MPLS procedures to propagate the AN failure towards the core network. This should involve removing associated routes (in access IGP case) and labels from its LIB and LFIB tables, and propagating the failure on the network side using BGP-LU and/or core IGP/LDP-DU procedures.

Upon AN coming back up, adjacent AN/AGN1x nodes automatically add routes pointing to recovered links based on the configured static routes or access IGP adjacency and link state updates. This should be then followed by LDP DoD label signaling and subsequent binding and installation of labels in LIB and LFIB tables.

3.5.3. AN/AGN Link Failure

Depending on the access topology and the failed link location different cases apply to the network operation after AN link failure (topology references from Section 2 in square brackets):

a. [all] - link failed, but at least one ECMP parallel link remains - nodes on both sides of the failed link must stop using the failed link immediately (local-repair), and keep using the remaining ECMP parallel links.

b. [I1, I, Y] - link failed, and there are no ECMP or alternative links and paths - nodes on both sides of the failed link must remove routes pointing to the failed link immediately from the RIB, remove associated labels from their LIB and LFIB tabels, and must send LDP label withdraw(s) to their upstream LSRs.

c. [U2, U, V, Y] - link failed, but at least one ECMP or alternate path remains - AN/AGN1x node must stop using the failed link and immediately switchover (local-repair) to the remaining ECMP path or alternate path. AN/AGN1x must remove affected next-hops and

labels from its tables and invoke LDP label withdraw as per point (a) above. If there is an AGN1x node terminating the failed link, it must remove routes pointing to the failed link immediately from the RIB, remove associated labels from their LIB and LFIB tabels, and must propagate the failure on the network side using BGP-LU and/or core IGP procedures.

If access IGP is used AN/AGN1x link failure will be propagated via IGP link updates across the access topology.

LDP DoD will also propagate the link failure by sending label withdraws to upstream AN/AGN1x nodes, and label release messages downstream AN/AGN1x nodes.

3.5.4. AGN Node Failure

AGN1x fails and all links to adjacent access nodes go down.

Depending on the access topology, following cases apply to the network operation after AGN1x node failure (topology references from Section 2 in square brackets):

a. [I1, I] - ANs are isolated from the network - AN adjacent to the failure must remove routes pointing to the failed AGN1x node immediately from the RIB, remove associated labels from their LIB and LFIB tabels, and must send LDP label withdraw(s) to their upstream LSRs. If access IGP is used, an IGP link update should be sent.

b. [U2, U, V, Y] - at least one ECMP or alternate path remains - AN adjacent to failed AGN1x must stop using the failed link and immediately switchover (local-repair) to the remaining ECMP path or alternate path. AN must remove affected routes and labels from its tables and invoke LDP label withdraw as per point (a) above.

Network side procedures for handling AGN1x node failure have been described in Seamless MPLS [I-D.ietf-mpls-seamless-mpls].

3.5.5. AGN Network-side Reachability Failure

AGN1x loses network reachability to a specific destination or set of network-side destinations.

In such event AGN1x must send LDP Label Withdraw messages to its upstream ANs, withdrawing labels for all affected /32 FECs. Upon receiving those messages ANs must remove those labels from their LIB and LFIB tables, and use alternative LSPs instead if available as

part of global-repair. In turn ANs should also sent Label Withdraw messages for affected /32 FECs to their upstream ANs.

If access IGP is used, and AGN1x gets completely isolated from the core network, it should stop advertising the default route 0/0 into the access IGP.

4. LDP DoD Procedures

Label Distribution Protocol is specified in [RFC5036], and all LDP Downstream-on-Demand implementations MUST follow this specification.

In the MPLS architecture [RFC3031], network traffic flows from upstream to downstream LSR. The use cases in this document rely on the downstream assignment of labels, where labels are assigned by the downstream LSR and signaled to the upstream LSR as shown in Figure 7.

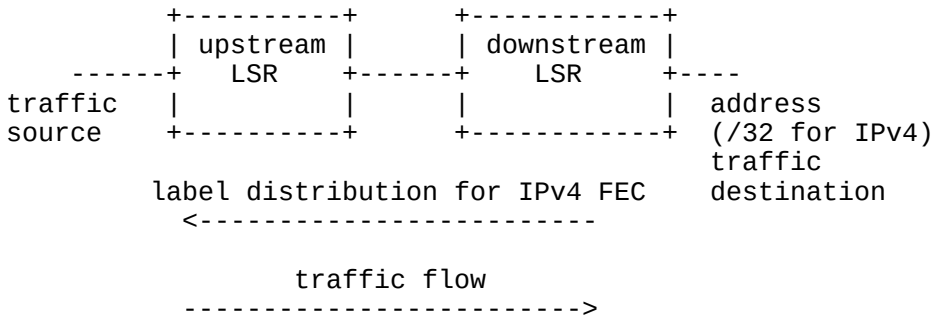


Figure 7: LDP label assignment direction

4.1. LDP Label Distribution Control and Retention Modes

LDP protocol specification [RFC5036] defines two modes for label distribution control, following the definitions in MPLS architecture [RFC3031]:

- Independent mode - an LSR recognizes a particular FEC and makes a decision to bind a label to the FEC independently from distributing that label binding to its label distribution peers. A new FEC is recognized whenever a new route becomes valid on the LSR.

- Ordered mode - an LSR binds a label to a particular FEC if it is the egress router for that FEC or if it has already received a label binding for that FEC from its next-hop LSR for that FEC.

Using independent label distribution control with LDP DoD and access static routing will prevent the access LSRs from propagating label binding failure along the access topology, making it impossible to switchover to an alternate path, even if such a path exists.

LDP protocol specification [RFC5036] defines two modes for label retention, following the definitions in MPLS architecture [RFC3031]:

- Liberal mode - LSR retains every label mappings received from a peer LSR, regardless of whether the peer LSR is the next-hop for the advertised mapping. This mode allows for quicker adaptation to routing changes.

- Conservative mode - LSR retains advertised label mappings only if they will be used to forward packets, that is only if they are received from a valid next-hop LSR according to routing. This mode allows LSR to maintain fewer labels, but slows down LSR adaptation to routing changes.

Using conservative label retention mode with LDP DoD will prevent the access LSRs (AN and AGN1x nodes) from implementing IPFRR LFA alternate based local-repair, as label mapping request can not be sent to alternate next-hops.

Adhering to the overall design goals of Seamless MPLS [I-D.ietf-mpls-seamless-mpls], specifically achieving a large network scale without compromising fast service restoration, all access LSRs (AN and AGN1x nodes) MUST use LDP DoD advertisement mode with:

- Ordered label distribution control - enables propagation of label binding failure within the access topology.

- Liberal label retention - enables pre-programming of alternate next-hops with associated FEC labels.

In Seamless MPLS [I-D.ietf-mpls-seamless-mpls] AGN1x node acts as an access ABR connecting access and metro domains. To enable failure propagation between those domains, access ABR MUST implement ordered label distribution control when redistributing access static routes and/or access IGP routes into the network-side BGP labeled unicast [RFC3107] or core IGP with LDP Downstream Unsolicited label advertisement.

4.2. IPv6 Support

Current LDP protocol specification [RFC5036] defines procedures and messages for exchanging FEC-label bindings over IPv4 and/or IPv6 networks. However number of IPv6 usage areas are not clearly

specified including: packet to LSP mapping for IPv6 destination router, no IPv6 specific LSP identifier, no LDP discovery using IPv6 multicast address, separate LSPs for IPv4 and IPv6, and others.

All of these issues and more are being addressed by [I-D.ietf-mpls-ldp-ipv6] that will update LDP protocol specification [RFC5036] in respect to the IPv6 usage. For the future deployment, LDP DoD use case and procedures described in this document SHOULD also support IPv6 for transport and services.

4.3. LDP DoD Session Negotiation

Access LSR/ABR should propose the Downstream-on-Demand label advertisement by setting "A" value to 1 in the Common Session Parameters TLV of the Initialization message. The rules for negotiating the label advertisement mode are specified in LDP protocol specification [RFC5036].

To establish a Downstream-on-Demand session between the two access LSR/ABRs, both should propose the Downstream-on-Demand label advertisement mode in the Initialization message. If the access LSR only supports LDP DoD and the access ABR proposes Downstream Unsolicited mode, the access LSR SHOULD send a Notification message with status "Session Rejected/Parameters Advertisement Mode" and then close the LDP session as specified in LDP protocol specification [RFC5036].

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

In case a PWE3 service is required between the adjacent access LSR/ ABR, and LDP DoD has been negotiated for IPv4 and IPv6 FECs, the same LDP session should be used for PWE3 FECs. Even if LDP 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 PWE3 LDP [RFC4447].

4.4. Label Request Procedures

4.4.1. Access LSR/ABR Label Request

Upstream access LSR/ABR will request label bindings from adjacent downstream access LSR/ABR based on the following trigger events:

a. Access LSR/ABR is configured with /32 static route with LDP DoD label request policy in line with intitial network setup use case described in Section 3.1.

b. Access LSR/ABR is configured with a service in line with service use cases described in Section 3.2 and Section 3.3.

c. Access LSR/ABR link to adjacent node comes up and LDP DoD session is established. In this case access LSR should send label request messages for all /32 static routes configured with LDP DoD policy and all /32 routes related to provisioned services that are not covered by default route. In line with use cases described in Section 3.5.

d. In all above cases requests MUST be sent to next-hop LSR(s) and alternate LSR(s).

Downstream access LSR/ABR will respond with label mapping message with a non-null label if any of the below conditions are met:

a. Downstream access LSR/ABR - requested FEC is an IGP or static route and there is an LDP label already learnt from the next-next-hop downstream LSR (by LDP DoD or LDP DU). If there is no label for the requested FEC and there is an LDP DoD session to the next-next-hop downstream LSR, downstream LSR MUST send a label request message for the same FEC to the next-next-hop downstream LSR. In such case downstream LSR will respond back to the requesting upstream access LSR only after getting a label from the next-next-hop downstream LSR peer.

b. Downstream access ABR only - requested FEC is a BGP labelled unicast route [RFC3107] and this BGP route is the best selected for this FEC.

Downstream access LSR/ABR may respond with 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. Label Request Retry

If an access LSR/ABR receives a "No route" Notification in response to its label request message, it should retry using an exponential backoff algorithm similar to the backoff algoritm mentioned in the LDP session negotiation described in Section 4.3.

If there is no response to the sent label request message, the LDP specification [RFC5036] (section A.1.1, page# 100) states that the

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 a Notification message.

Thus, if there is no response from the downstream peer, the access LSR/ABR should not send a duplicate label request message again.

If the static route corresponding to the FEC gets deleted or if the DoD request policy is modified to reject the FEC before receiving the label mapping message, then the access LSR/ABR should send a Label Abort message to the downstream LSR.

4.4.3. Label Request with Fast-Up Convergence

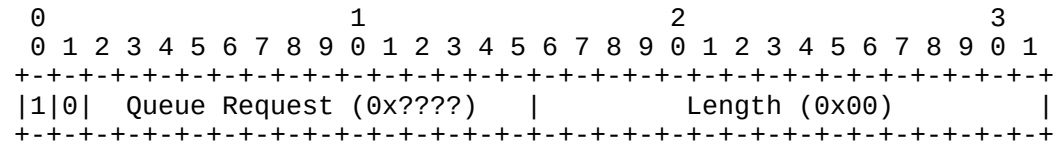
In some conditions, the exponential backoff algorithm usage described in Section 4.4.2 may result in a longer than desired wait time to get a successful LDP label to route mapping. An example is when a specific route is unavailable on the downstream LSR when the label mapping request from the upstream is received, but later comes back. In such case using the exponential backoff algorithm may result in a max delay wait time before the upstream LSR sends another LDP label request.

Fast-up convergence can be addressed with a minor extension to the LDP DoD procedure, as described in this section. The downstream and upstream LSRs SHOULD implement this extension if up convergence improvement is desired.

The extension consists of the upstream LSR indicating to the downstream LSR that the label request should be queued on the downstream LSR until the requested route is available.

To implement this behavior, a new Optional Parameter is defined for use in the Label Request message:

Optional Parameter	Length	Value
Queue Request TLV	0	see below



U-bit = 1
Unknown TLV bit is set to 1. If this optional TLV is unknown, it should be ignored without sending "no route" notification. Ensures backward compatibility.

F-bit = 0
Forward unknown TLV bit is set to 0. The unknown TLV is not forwarded.

Type
Queue Request Type value to be allocated by IANA.

Length = 0x00
Specifies the length of the Value field in octets.

The operation is as follows.

To benefit from the fast-up convergence improvement, the upstream LSR sends a Label Request message with a Queue Request TLV.

If the downstream LSR supports the Queue Request TLV, it verifies if route is available and if so it replies with label mapping as per existing LDP procedures.

If the route is not available, the downstream LSR queues the request and replies as soon as the route becomes available. In the meantime, it does not send a "no route" notification back. When sending a label request with the Queue Request TLV, the upstream LSR does not retry the Label Request message if it does not receive a reply from its downstream peer

If the upstream LSR wants to abort an outstanding label request while the Label Request is queued in the downstream LSR, the upstream LSR sends a Label Abort Request message, making the downstream LSR to remove the original request from the queue and send back a notification Label Request Aborted [RFC5036].

If the downstream LSR does not support the Queue Request TLV, it will silently ignores it, and sends a "no route" notification back. In this case the upstream LSR invokes the exponential backoff algorithm described in Section 4.4.2.

This described procedure ensures backward compatitibility.

4.5. Label Withdraw

If an MPLS label on the downstream access LSR/ABR is no longer valid, the downstream access LSR/ABR withdraws this FEC/label binding from the upstream access LSR/ABR with the Label Withdraw Message [RFC5036] with a specified label TLV or with an empty label TLV.

Downstream access LSR/ABR SHOULD withdraw a label for specific FEC in the following cases:

a. If LDP DoD ingress label is associated with an outgoing label assigned by BGP labelled unicast route, and this route is withdrawn.

b. If LDP DoD ingress label is associated with an outgoing label assigned by LDP (DoD or DU) and the IGP route is withdrawn from the RIB or downstream LDP session is lost.

c. If LDP DoD ingress label is associated with an outgoing label assigned by LDP (DoD or DU) and the outgoing label is withdrawn by the downstream LSR.

d. If LDP DoD ingress label is associated with an outgoing label assigned by LDP (DoD or DU), route next-hop changed and

   * there is no LDP session to the new next-hop. To minimize probability of this, the access LSR/ABR should implement LDP-IGP synchronization procedures as specified in [RFC5443].

   * there is an LDP session but no label from downstream LSR. See note below.

e. If access LSR/ABR is configured with a policy to reject exporting label mappings to upstream LSR.

The upstream access LSR/ABR responds to the Label Withdraw Message with the Label Release Message [RFC5036].

After sending label release message to downstream access LSR/ABR, the upstream access LSR/ABR should resend label request message, assuming upstream access LSR/ABR still requires the label.

Downstream access LSR/ABR should withdraw a label if the local route configuration (e.g. /32 loopback) is deleted.

Note: For any events inducing next hop change, downstream access LSR/

ABR should attempt to converge the LSP locally before withdrawing the label from an upstream access LSR/ABR. For example if the next-hop changes for a particular FEC and if the new next-hop allocates labels by LDP DoD session, then the downstream access LSR/ABR must send a label request on the new next-hop session. If downstream access LSR/ABR doesn't get label mapping for some duration, then and only then downstream access LSR/ABR must withdraw the upstream label.

4.6. Label Release

If an access LSR/ABR does not need any longer a label for a FEC, it sends a Label Release Message [RFC5036] to the downstream access LSR/ABR with or without the label TLV.

If upstream access LSR/ABR receives an unsolicited label mapping on DoD session, they should release the label by sending label release message.

Access LSR/ABR should send a label release message to the downstream LSR in the following cases:

a. If it receives a label withdraw from the downstream access LSR/ABR.

b. If the /32 static route with LDP DoD label request policy is deleted.

c. If the service gets decommissioned and there is no corresponding /32 static route with LDP DoD label request policy configured.

d. If the route next-hop changed, and the label does not point to the best or alternate next-hop.

e. If it receives a label withdraw from a downstream DoD session.

4.7. Local Repair

To support local-repair with ECMP and IPFRR LFA, access LSR/ABR MUST request labels on both best next-hop and alternate next-hop LDP DoD sessions as specified in the label request procedures in Section 4.4. This will enable access LSR/ABR to pre-program the alternate forwarding path with the alternate label(s), and invoke IPFRR LFA switch-over procedure if the primary next-hop link fails.

5. IANA Considerations

5.1. LDP TLV TYPE

This document uses a new a new Optional Parameter Queue Request TLV in the Label Request message defined in Section 4.4.3. IANA already maintains a registry of name LDP "TLV TYPE NAME SPACE" defined by RFC5036. The following value is suggested for assignment:

TLV type	Description
0x0971	Queue Request TLV

6. Security Considerations

7. Acknowledgements

The authors would like to thank Nischal Sheth, Nitin Bahadur, Nicolai Leymann and Ina Minei 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.

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

8.2. Informative References

[I-D.ietf-mpls-ldp-ipv6] Asati, R., Manral, V., Papneja, R., and C. Pignataro, "Updates to LDP for IPv6", draft-ietf-mpls-ldp-ipv6-05 (work in progress), August 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.

[RFC4364] Rosen, E. and Y. Rekhter, "BGP/MPLS IP Virtual Private Networks (VPNs)", RFC 4364, February 2006.

[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.

[RFC5283] Decraene, B., Le Roux, JL., 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.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:

Luca Martini
Cisco Systems, Inc.
9155 East Nichols Avenue, Suite 400
Englewood, CO 80112
USA

Phone:
Fax:
Email: lmartini@cisco.com
URI:

MPLS Working Group
Internet Draft
Intended status: Standards Track
Expires: April 24, 2012

Nabil Bitar
Verizon

Himanshu Shah
Ciena

George Swallow
Cisco

October 24, 2011

ISIS MPLS Explicit NULL Label
draft-bitar-mpls-isis-explicit-null-label-00.txt

Abstract

There is need to support IP interfaces on the top of GMPLS packet Label Switched Paths (LSPs), and enable IP routing (e.g., OSPF-TE, ISIS-TE) and MPLS protocols on these interfaces. Traffic on an IP/MPLS interface can be user traffic or control traffic. In addition, it can be MPLS, IP or ISIS. Multiplexing IP and MPLS packets over the same LSP is supported in the current MPLS architecture. However, multiplexing IP, MPLS, and ISIS packets over the same LSP is not currently supported. This draft proposes the definition of an explicit ISIS NULL label to enable this type of multiplexing to take place.

Table of Contents

1. Introduction

RFC 4206 [RFC 4206] defines the concept of a forwarding adjacency (FA) built on a Traffic Engineered (TE) LSP. An FA can be unidirectional and signaled via RSVP-TE, or bidirectional and signaled via RSVP-TE with GMPLS extensions. It is signaled over the same network on which it is routed. An FA is included in the network IGP link state database as a TE link but used only for forwarding. That is, it is never used to establish a routing adjacency. Routing adjacencies are necessary in a link-state IGP network for topology discovery and link-state information dissemination. FAs capitalize on the existence of routing adjacencies, and routers make use of the topology information exchanged over these adjacencies to establish the FAs.

In MPLS-TP, client network islands that belong to the same IGP are interconnected over an MPLS-TP [RFC 5921] server network in an overlay model. A client network island is connected to the MPLS-TP network via a border client router. Two client border-routers need to form a routing adjacency over a GMPLS LSP signaled through the MPLS-TP network via GMPLS UNI signaling. The GMPLS UNI is the interface between the client border node and the connected MPLS-TP network edge.

This document introduces the explicit ISIS NULL label that enables the establishment of an ISIS(-TE) routing adjacency over a GMPLS LSP, and to treat that routing adjacency as any IP adjacency, enabling MPLS signaling, IP and MPLS multicast signaling/routing, and MPLS and IP forwarding over that adjacency.

2. Operational Procedures

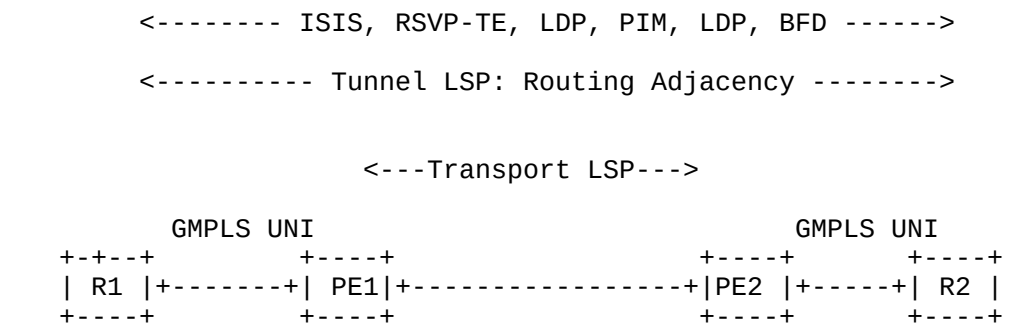


Figure 1: Reference Model for GMPLS LSP Tunnel as an IP interface

Figure 1 depicts a reference model for the GMPLS UNI and GMPLS LSP routing adjacency, referred to as tunnel LSP. R1 connects to the MPLS transport network via a GMPLS UNI interface between R1 and PE1. R2 connects to the MPLS transport network via a GMPLS UNI interface between R2 and PE2. A GMPLS tunnel LSP is signaled over the GMPLS

UNI and across the MPLS transport network. The LSP endpoint at R1 is presented as an IP interface to R1. The LSP endpoint at R2 is presented as an IP interface to R2. ISIS(-TE) is enabled on these IP interfaces. In addition, other IP-based protocols such as RSVP-TE, LDP, PIM-SM(SSM), etc., can be enabled on these interfaces. It should be noted that if OSPF is enabled in addition or instead of ISIS over these interfaces, OSPF packets can be carried over the GMPLS LSP tunnel using the existing MPLS encapsulation architecture [RFC 3032] for transporting IP packets.

When the tunnel LSP is active at R1 and R2, ISIS adjacency formation starts and ISIS adjacency is established. Subsequent to that ISIS Link state packets are flooded over that LSP as on any other ISIS link. Other LSPs can be established over this ISIS link (the LSP tunnel) via RSVP-TE, GMPLS, LDP, etc.

When R1 sends an ISIS packet to R2, it first imposes the explicit ISIS NULL label whose value TBD, with the S-bit to 1, the TC set to a configured value, and TTL set to 1, and then encapsulates the MPLS packet with the LSP tunnel header.

When R1 sends R2 an IPv4 control protocol (e.g., RSVP-TE, LDP, PIM), or a user IPv4 packet, it encapsulates the IPv4 packet with the LSP tunnel header.

When R1 sends R2 an IPv6 control protocol (e.g., RSVP-TE, LDP, PIM), or a user IPv6 packet, it first imposes on the packet the IPv6 Explicit NULL label (label value = 2) with the S bit set to 1, followed by a label header corresponding to the GMPLS LSP tunnel.

When R1 sends an MPLS packet to R2, it encapsulates the MPLS packet with the LSP tunnel header. In this case, R1 may be a transit node for the LSP whose MPLS packet is being switched across the tunnel GMPLS LSP, or the head-end of that LSP.

Upon receiving a packet over the GMPLS LSP tunnel configured as a routing adjacency, R1 performs the following processing:

1. It pops the GMPLS LSP label, preserving the context of that label as an IP interface

2. If the encapsulated packet is an MPLS packet, as indicated by the outer label (tunnel label) tag S-bit set to 0, R1 performs a label lookup on the top label, after popping the tunnel label. The following cases exist:

   a. The label matches the ISIS Explicit NULL label value: the encapsulated packet is an ISIS packet. The ISIS packet is sent to the control plane.

   b. The label matches the IPv6 Explicit NULL label value: the encapsulated packet is an IPv6 packet, the IPv6 Explicit Null label is popped, and the IPv6 packet is routed appropriately.

   c. Otherwise, the label is switched, or popped depending on the label context.

3. If the encapsulated packet is not an MPLS packet, (i.e., the tunnel label had the S bit set to 1), it is assumed that the encapsulated packet is an IPv4 packet.

3. ISIS Packet Encapsulation format

Figure 2 depicts the protocol stack for an ISIS packet encapsulated over the GMPLS LSP tunnel.

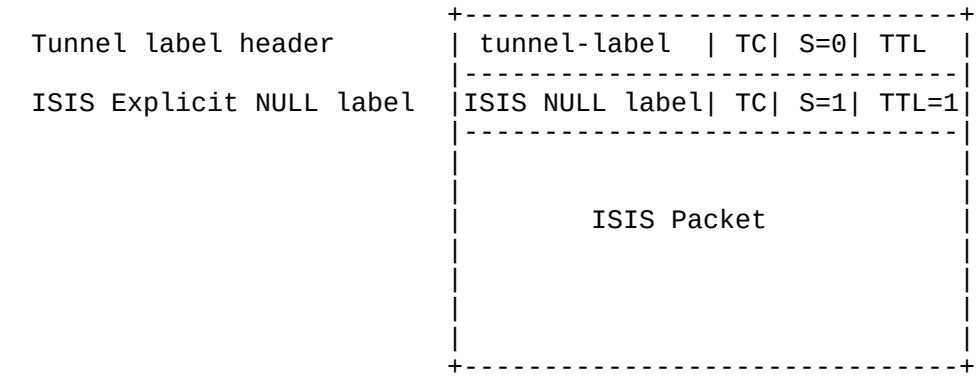


Figure 2: Encapsulation of an ISIS packet in a tunnel LSP treated as a routing adjacency

4. IANA Consideration

This document requires the designation of a label value in the reserved MPLS Label space for the ISIS Explicit NULL label.

5. Security Considerations

No new security issues are introduced in this document beyond what is addressed for MPLS, GMPLS and all the IP protocols.

6. References

6.1. Normative References

[RFC 3032] Rosen, E., et. al, "MPLS Label Stack Encoding", RFC 3032, January, 2011.

[RFC 5921] Bocci, M., et. al, "A Framework for MPLS in Transport Networks", RFC 5921, July 2010.

6.2. Informative References

[RFC 4206] Kompella, K., and Rekhter, Y., "Label Switched Paths (LSP) Hierarchy with Generalized Multi-Protocol Label Switching (GMPLS) Traffic Engineering (TE)", RFC 4206, October 2005.

Authors' Addresses

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

Himanshu Shah
Ciena Corp
Email: hshah@ciena.com

George Swallow
Cisco
Email: swallow@cisco.com

Network Working Group
Internet-Draft
Updates: 4379 (if approved)
Intended status: Standards Track
Expires: March 31, 2012

M. Chen
Huawei Technologies Co., Ltd
P. Pan
Infinera
C. Pignataro
R. Asati
Cisco

September 28, 2011

Label Switched Path (LSP) Ping for IPv6 Pseudowire FECs
draft-chen-mpls-ipv6-pw-lsp-ping-02

Abstract

Multi-Protocol Label Switching (MPLS) Label Switched Path (LSP) Ping and Traceroute mechanisms are commonly used to detect and isolate data plane failures in all MPLS LSPs including Pseudowire (PW) LSPs. The PW LSP Ping and Traceroute elements, however, are not specified for IPv6 address usage.

Specifically, the Pseudowire FEC sub-TLVs for the Target FEC Stack in the LSP Ping and Traceroute mechanism are implicitly defined only for IPv4 Provider Edge (PEs) routers, and are not applicable for the case where PEs use IPv6 addresses. There is, additionally, a degree of potential ambiguity in the specification of these sub-TLVs since the address family is not explicitly specified but it is to be inferred from the sub-TLV length.

This document updates RFC4379 to explicitly constraint these existing PW FEC sub-TLVs for IPv4 LDP sessions, and extends Pseudowire LSP Ping to the IPv6 scenario where an IPv6 LDP session is used to signal the Pseudowire (i.e., where the Sender's and Receiver's IP addresses are IPv6 addresses.) This is done by defining two new LSP Ping sub-TLVs for IPv6 Pseudowire FECs.

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 March 31, 2012.

Copyright Notice

Table of Contents

1. Introduction

Multi-Protocol Label Switching (MPLS) Label Switched Path (LSP) Ping and Traceroute are defined in [RFC4379]. This mechanism can be used to detect and isolate data plane failures in all MPLS Label Switched Paths (LSPs) including Pseudowires (PWs). Currently, three PW related Target Forwarding Equivalence Class (FEC) sub-TLVs (FEC 128 Pseudowire-Deprecated, FEC 128 Pseudowire-Current, and FEC 129 Pseudowire) are defined (see Section 3.2 of [RFC4379]). These sub-TLVs contain the source and destination addresses of the target LDP session, and currently only IPv4 target LDP session is covered. Despite the fact that the IP address family is not explicit in the sub-TLV definition, this can be inferred indirectly only calculating the Length of the sub-TLVs. When IPv6 target LDP session is used, these existing sub-TLVs can not therefore be used. Additionally, all other sub-TLVs are defined in pairs, one for IPv4 and another for IPv6, and not for PW sub-TLVs.

This document updates [RFC4379] to make explicit the IPv4 nature of the existing PW sub-TLVs, and also defines two new Target FEC sub-TLVs (IPv6 FEC 128 Pseudowire sub-TLV and IPv6 FEC 129 Pseudowire sub-TLV) to extend the application of PW LSP Ping and Traceroute to the IPv6 usage when an IPv6 LDP session [I-D.ietf-mpls-ldp-ipv6] is used to signal the Pseudowire. Note that FEC 128 Pseudowire (Deprecated) is not defined for IPv6 in this document.

2. IPv4 Pseudowire Sub-TLVs

This document updates Section 3.2 and Sections 3.2.8 through 3.2.10 of [RFC4379] as follows and as indicated in Section 4 and Section 6. This is done to avoid any potential ambiguity, confusion, and backwards compatibility issues.

Sections 3.2.8 through 3.2.10 of [RFC4379] list the PW sub-TLVs and state:

"FEC 128" Pseudowire (Deprecated)

"FEC 128" Pseudowire

"FEC 129" Pseudowire

These names and titles are now changed to:

IPv4 "FEC 128" Pseudowire (Deprecated)

IPv4 "FEC 128" Pseudowire

IPv4 "FEC 129" Pseudowire

Additionally, when referring to the PE addresses, these three sections state:

Sender’s PE Address

Remote PE Address

These are now updated to say:

Sender’s PE IPv4 Address

Remote PE IPv4 Address

3. IPv6 Pseudowire Sub-TLVs

3.1. IPv6 FEC 128 Pseudowire Sub-TLV

IPv6 FEC 128 Pseudowire sub-TLV has the consistent structure with FEC 128 Pseudowire sub-TLV as described in Section 3.2.9 of [RFC4379]. The encoding of IPv6 FEC 128 Pseudowire sub-TLV is as follows:

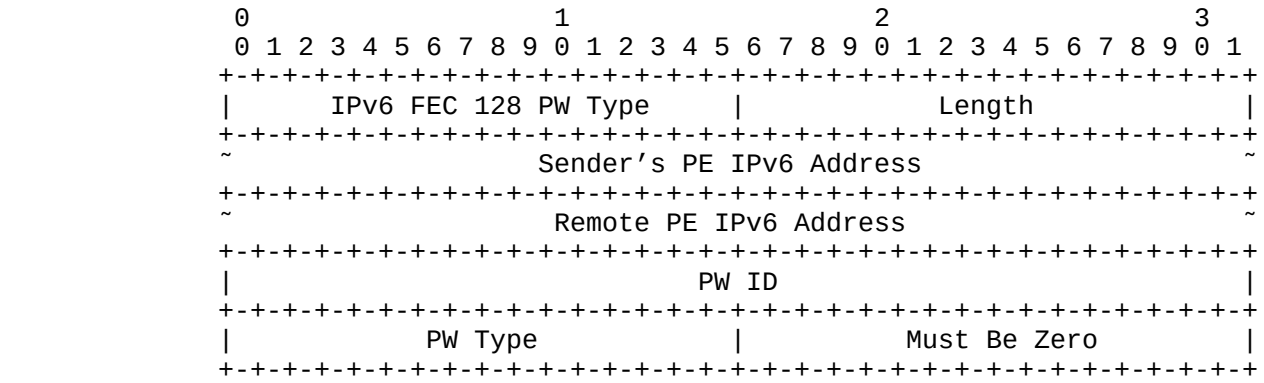


Figure 1: IPv6 FEC 128 Pseudowire

IPv6 FEC 128 PW: TBD.

Length: it defines the length in octets of the value field of the sub-TLV and its value is 38.

Sender's PE IPv6 Address: The source IP address of the target IPv6 LDP session.

Remote PE IPv6 Address: The destination IP address of the target IPv6 LDP session.

PW ID: Same as FEC 128 Pseudowire [RFC4379].

PW Type: Same as FEC 128 Pseudowire [RFC4379].

3.2. IPv6 FEC 129 Pseudowire Sub-TLV

IPv6 FEC 129 Pseudowire sub-TLV has the consistent structure with FEC 129 Pseudowire sub-TLV as described in Section 3.2.10 of [RFC4379]. The encoding of IPv6 FEC 129 Pseudowire is as follows:

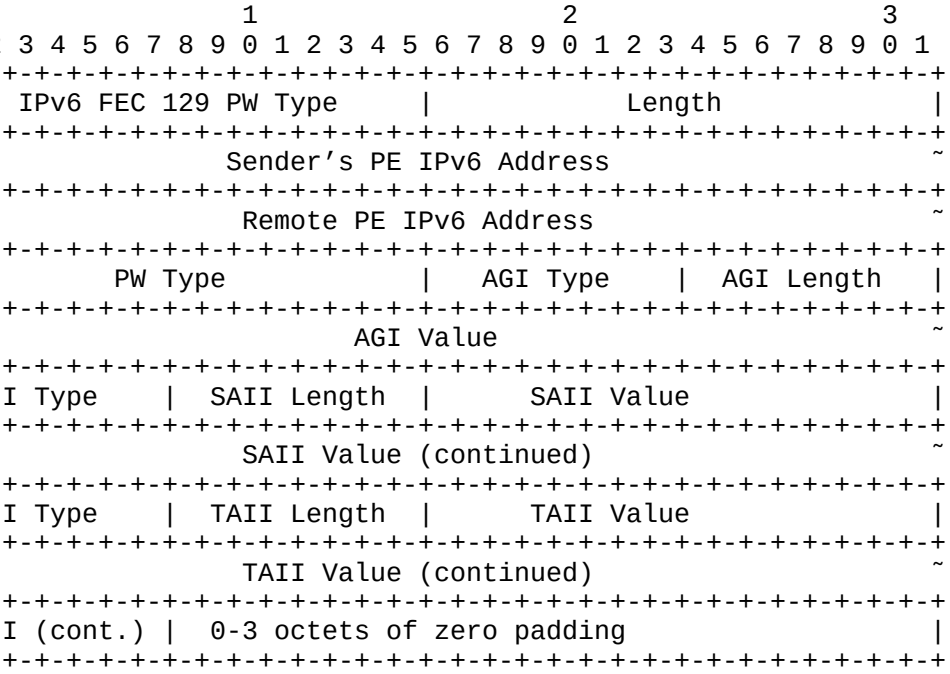


Figure 2: IPv6 FEC 129 Pseudowire

IPv6 FEC 129 PW: TBD.

The Length of this TLV is 40 + AGI length + SAII length + TAII length. Padding is used to make the total length a multiple of 4; the length of the padding is not included in the Length field.

Sender's PE IPv6 Address: The source IP address of the target IPv6 LDP session.

Remote PE IPv6 Address: The destination IP address of the target IPv6 LDP session.

The other fields are same as FEC 129 Pseudowire [RFC4379].

4. Summary of Changes

Section 3.2 of [RFC4379] tabulates all the sub-TLVs for the Target FEC Stack. Per the change described in Section 2 and Section 3, the table would show the following:

Sub-Type	Length	Value Field
...		
9	10	IPv4 "FEC 128" Pseudowire (deprecated)
10	14	IPv4 "FEC 128" Pseudowire
11	16+	IPv4 "FEC 129" Pseudowire
...		
TBD	38	IPv6 "FEC 128" Pseudowire
TBD	40+	IPv6 "FEC 129" Pseudowire

5. Operation

This document does not define any new procedures. The process described in [RFC4379] MUST be used.

6. IANA Considerations

IANA is requested to perform the following assignments in the "Multi-Protocol Label Switching (MPLS) Label Switched Paths (LSPs) Ping Parameters" registry, "TLVs and sub-TLVs" sub-registry.

[RFC Editor: To be REMOVED prior to publication. This registration should take place at <http://www.iana.org/assignments/mpls-lsp-ping-parameters/mpls-lsp-ping-parameters.xml#mpls-lsp-ping-parameters-7>]

Update the Value fields of these three Sub-TLVs, adding the "IPv4" qualifier (see Section 2), and update the Reference to point to this document:

Type	Sub-Type	Value Field
1	9	IPv4 "FEC 128" Pseudowire (Deprecated)
1	10	IPv4 "FEC 128" Pseudowire
1	11	IPv4 "FEC 129" Pseudowire

Create two new entries for the Sub-Type field of Target FEC TLV (see Section 3):

Type	Sub-Type	Value Field
1	TBD1	IPv6 "FEC 128" Pseudowire
1	TBD2	IPv6 "FEC 129" Pseudowire

7. Security Considerations

This draft does not introduce any new security issues, the security mechanisms defined in [RFC4379] apply here.

8. Acknowledgements

The authors gratefully acknowledge review and comments of Vanson Lim and Tom Petch.

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.

[RFC4379] Kompella, K. and G. Swallow, "Detecting Multi-Protocol Label Switched (MPLS) Data Plane Failures", RFC 4379, February 2006.

9.2. Informative References

[I-D.ietf-mpls-ldp-ipv6] Asati, R., Manral, V., Papneja, R., and C. Pignataro, "Updates to LDP for IPv6", draft-ietf-mpls-ldp-ipv6-05 (work in progress), August 2011.

Authors' Addresses

Mach(Guoyi) Chen
Huawei Technologies Co., Ltd
No. 3 Xinxi Road, Shang-di, Hai-dian District
Beijing 100085
China

Email: mach@huawei.com

Ping Pan
Infinera
US

Email: ppan@infinera.com

Carlos Pignataro
Cisco Systems
7200-12 Kit Creek Road
Research Triangle Park, NC 27709
US

Email: cpignata@cisco.com

Rajiv Asati
Cisco Systems
7025-6 Kit Creek Road
Research Triangle Park, NC 27709
US

Email: rajiva@cisco.com

Internet Engineering Task Force
Internet-Draft
Intended status: Standards Track
Expires: May 2, 2012

H. Chen
Huawei Technologies
N. So
Verizon Inc.
A. Liu
Ericsson

October 30, 2011

Extensions to RSVP-TE for P2MP LSP Egress Local Protection
draft-chen-mpls-p2mp-egress-protection-04.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 May 2, 2012.

Copyright Notice

Table of Contents

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.

4.1. An Example of Egress Local Protection

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

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

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

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

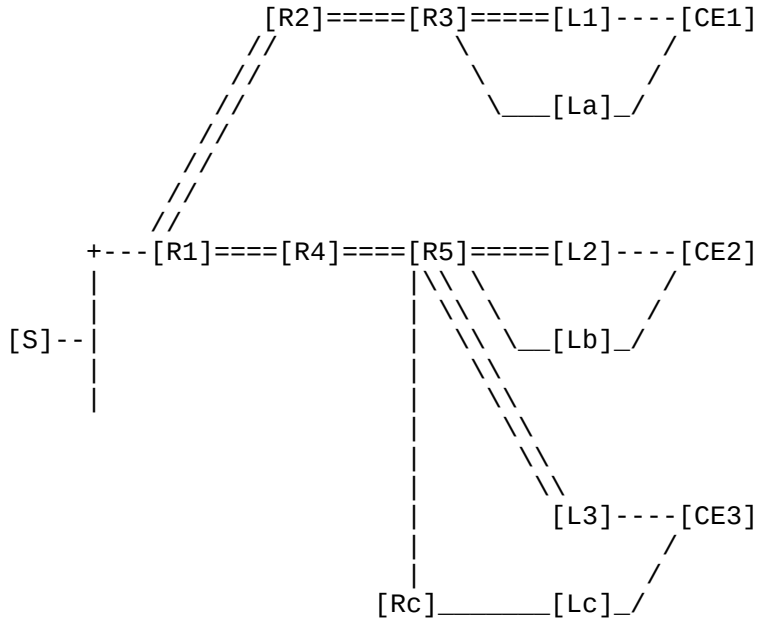


Figure 1: P2MP sub LSP for Locally Protecting Egress

4.2. Set up of Backup sub LSP

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

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

4.3. Forwarding State for Backup sub LSP(s)

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

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

4.4. Detection of Egress Node Failure

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

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

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

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

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

5. Egress Local Protection with FRR

RFC4875 "Extensions to RSVP-TE for P2MP TE LSPs" describes how to use the existing FRR to locally protect failures in a link or intermediate node of a P2MP LSP. Thus, all the links, the egress nodes and the intermediate nodes of a P2MP LSP can be locally protected through using the egress local protection mechanism described above and the FRR.

All the egress nodes of the P2MP LSP can be locally protected through using the egress local protection. All the links and the

intermediate nodes of the LSP can be locally protected by using the FRR.

All the intermediate nodes of the P2MP LSP may be locally protected by some other methods such as the method proposed in "P2MP MPLS-TE Fast Reroute with P2MP Bypass Tunnels". Note that the methods for locally protecting all the links and the intermediate nodes of a P2MP LSP are out of scope of this document.

6. Representation of a Backup Sub LSP

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

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

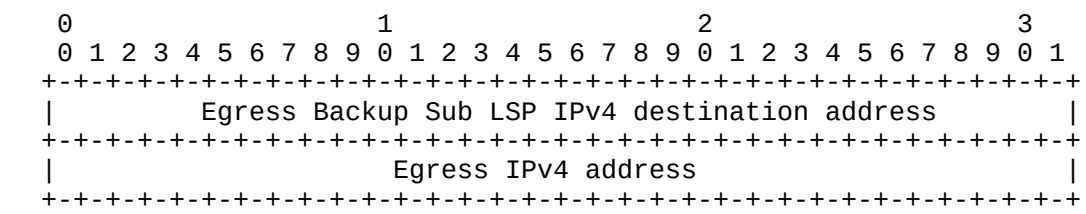
6.1. EGRESS_BACKUP_SUB_LSP Object

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

6.1.1. EGRESS_BACKUP_SUB_LSP IPv4 Object

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

EGRESS_BACKUP_SUB_LSP Class = 50,
EGRESS_BACKUP_SUB_LSP_IPv4 C-Type = 3

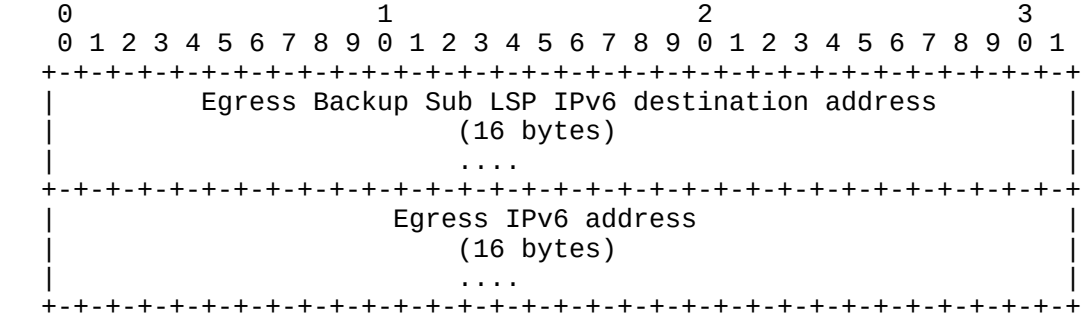


Egress Backup Sub LSP IPv4 destination address
IPv4 address of the backup sub LSP destination is the backup egress node.
Egress IPv4 address
IPv4 address of the egress node

6.1.2. EGRESS_BACKUP_SUB_LSP IPv6 Object

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

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

6.2. 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 that of 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.

7. 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 a primary egress node of a P2MP LSP through including an egress backup sub LSP descriptor list.

7.1. 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> ]
```

7.2. Processing of Path Message

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

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

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

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

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

8. IANA Considerations

TBD

9. Acknowledgement

The authors would like to thank Richard Li, Neil Harrison, Lei Liu, Kannan Sampath, Yimin Shen and Quintin Zhao for their valuable comments and suggestions 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.

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

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

10.2. 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

Autumn Liu
Ericsson
CA
USA

Email: autumn.liu@ericsson.com

Internet Engineering Task Force
Internet-Draft
Intended status: Standards Track
Expires: May 2, 2012

H. Chen
Huawei Technologies
N. So
Verizon Inc.
A. Liu
Ericsson

October 30, 2011

Extensions to RSVP-TE for P2MP LSP Ingress Local Protection
draft-chen-mpls-p2mp-ingress-protection-04.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 May 2, 2012.

Copyright Notice

Table of Contents

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.

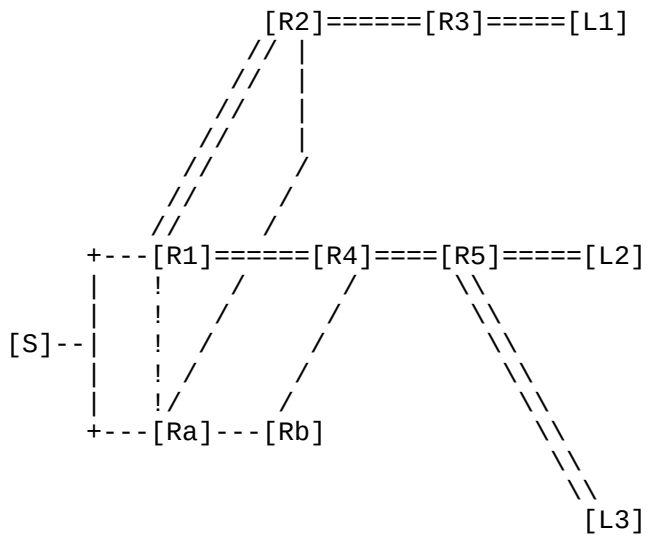


Figure 1: P2MP sub Tree for Locally Protecting Ingress

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 ingress node sends the P2MP LSP information to the backup ingress node. The backup ingress node initiates the creation of the backup P2MP sub tree from itself to the next-hop nodes of the ingress node.

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

4.3. Forwarding State for Backup P2MP sub Tree

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

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

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

4.4. Detection of Failure around Ingress

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

- The failure of the ingress node (e.g. R1 of figure 1).
- 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 over the link and to the backup ingress via the ingress.

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

5. Ingress Local Protection with FRR

RFC4875 "Extensions to RSVP-TE for P2MP TE LSPs" describes how to use the existing FRR to locally protect failures in a link or intermediate node of a P2MP LSP. Thus, the ingress node , all the

links and the intermediate nodes of a P2MP LSP can be locally protected through using the ingress local protection mechanism described above and the FRR.

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

All the intermediate nodes of the P2MP LSP may be locally protected by some other methods such as the method proposed in "P2MP MPLS-TE Fast Reroute with P2MP Bypass Tunnels". Note that the methods for locally protecting all the links and the intermediate nodes of a P2MP LSP are out of scope of this document.

6. LSP Information Message

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

6.1. Format of LSP Information Message

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

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

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

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

6.2. Processing of LSP Information Message

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

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

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

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

as follows.

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

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

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

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

The 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 authors would like to thank Richard Li, Rahul Aggarwal, Neil Harrison, Lei Liu, Kannan Sampath, Yimin Shen and Quintin Zhao for their valuable comments and suggestions 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.

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

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

Autumn Liu
Ericsson
CA
USA

Email: autumn.liu@ericsson.com

MPLS Working Group
Internet-Draft
Intended status: Standards Track
Expires: May 3, 2012

T. Cheung
J. Ryoo
ETRI
Y. Weingarten
N. Sprecher
Nokia Siemens Networks
D. King
Old Dog Consulting

October 31, 2011

MPLS-TP Shared Mesh Protection
draft-cheung-mpls-tp-mesh-protection-04.txt

Abstract

This document describes a mechanism to address the requirement to support 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 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.

Table of Contents

1. Introduction

The MPLS Transport Profile (MPLS-TP) is a packet transport technology based on a profile of the MPLS and Pseudowires (PW) as described in [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 protect 100% of the traffic on the protected path (Requirement 58).

1+1 and 1:1 linear protection meets these requirements by reserving the equivalent amount of network resources for the protection paths as is allocated to the normal traffic that is being 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 between the same two end-points. This improves overall network utilization, but the resource (bandwidth) allocated to a protection path is typically not sufficient to protect multiple simultaneous failures on different working paths. If multiple working paths require concurrent protection switching, the path with the highest priority should be protected as described in [RFC6372].

In 1+1 and 1:1 protection, the end nodes of the working path must be the same as those of the protection path. Similarly in 1:n protection 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 type of protection can make very efficient use of network resources, but requires coordination of several segments in order to ensure that only a single traffic flow is switched to the protection resources at any time.

[RFC4428] defines two shared mesh recovery schemes named (1:1)^n and (M:N)^n. The (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 any time. In this case, it achieves network utilization similar to 1:n protection.

It should be noted that the (1:1)^n protection scheme described 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 the same end nodes.

This document defines a data-plane shared mesh protection mechanism based on the concept of the (1:1)^n recovery scheme described in [RFC4428] and a protocol for coordination of the shared protection resources. The actual protection switching is controlled by end-to-end linear protection, while the usage of the shared resources is based on the protection switching priority assigned to each pair of working and protection paths.

The shared mesh protection mechanism defined in this document utilizes the existing MPLS-TP linear protection switching mechanism, 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 the 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. However, the intermediate nodes that are part of the shared segments need to coordinate the resource allocation on the shared nodes and this coordination will be addressed by the protocol proposed in this document.

2. Conventions Used in this Document

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

2.1. Acronyms

This draft uses the following acronyms:

G-ACh	Generic Associated Channel Header
LoP	Lockout of Protection
LP	Linear Protection
LSP	Label Switched Path
MIP	Maintenance Entity Group Intermediate Point
MPLS-TP	Transport Profile for MPLS
P2P	Point-to-point
P2MP	Point-to-multipoint
PW	Pseudowire
SEN	Shared End Node
SMP	Shared Mesh Protection
SMPG	Shared Mesh Protection Group
SPME	Sub-Path Maintenance Entity
SSN	Shared Start Node

2.2. Definitions and Terminology

This document defines two protection domains as follows:

- End-to-end linear protection domain: A protection domain as defined in [RFC6372] for protecting a P2P or P2MP LSP. It consists of two or more end points at the boundary of the domain and a working path and a protection path between the end nodes. An end-to-end linear protection switching protocol runs within the domain.

- 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.

In addition, we define the following:

- o Shared mesh protection group (SMPG): a protection group includes the pairs of working and protection paths, whose working paths do not belong to a single SRLG and whose protection paths share a single sub-segment. Note that an LSP may belong to multiple protection groups.

3. Shared Mesh Protection Architecture

The shared mesh protection domain shown in Figure 1 has two end-to-end linear protection domains. One consists of the two end nodes A and E and includes one working path, ABCDE, and one dedicated protection path APQRE. The second consists of end nodes V and Z and one working path, VWXYZ, and the dedicated protection path, VPQRZ. Those two domains share a common segment PQR for their protection path. This illustrates a simple configuration of shared mesh protection. Note that the two working paths, ABCDE and VWXYZ, do not share end points so they cannot make use of 1:n protection even though they also do not share any potential common points of failure.

It is possible to apply linear protection to each of these working paths individually. If there are no failures affecting either of the two working paths, the network segment PQR carries no traffic (or only interruptible extra traffic). In the event of only one failure, the segment PQR carries traffic from the working path that detected the failure. Only in the event that there are failures detected on both of the working paths is there a conflict over the appropriate use of the shared PQR segment. It is important to note that there are two distinct LSPs (i.e. APQRE and VPQRZ) that are signaled over the shared segment, and that although we refer to the singular segment, the traffic is actually being transported on separated transport paths.

Thus, it is possible for the network resources of 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 need to be reserved to provide protection of the multiple paths within the same protection group.

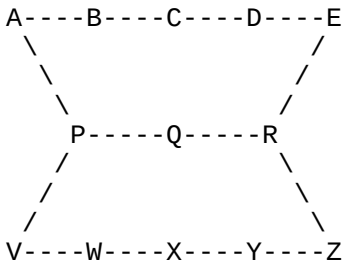


Figure 1: A Shared Mesh Protection Topology

3.1. Shared Mesh Protection Group

The two working paths in Figure 1, ABCDE and VWXYZ, are considered a Shared Mesh Protection Group (SMPG). Such a group is defined as the set of working paths whose protection path share the resources of a single shared segment. As pointed out above, there are individual protection LSP for each of the LP domains, however the resources that are being shared are the nodes, ports, links and bandwidth of the segment.

The shared resources, for example bandwidth capacity, should be reserved in partitions according to the different SMPGs at the particular segment.

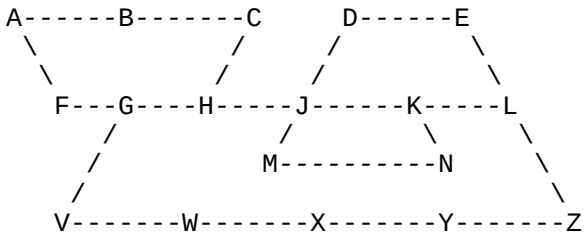


Figure 2: Shared Mesh Protection Groups

To further clarify, consider the mesh network in Figure 2. In this figure we have the following working paths and corresponding protection paths:

Wx	working path	protection path
W1	A-B-C	A-F-G-H-C
W2	D-E	D-J-K-L-E
W3	M-N	M-J-K-N
W4	V-W-X-Y-Z	V-G-H-J-K-L-Z

In this network we would define three SMPG - characterized by the three shared segments -

1. S1 segment G-H - shared by W1 and W4

2. S2 segment J-K - shared by W2, W3, and W4

3. S3 segment K-L - shared by W2 and W4

The shared segment is always the smallest segment that is shared by multiple protection paths. Therefore, even though segment J-K-L is shared by W2 and W4, we split this into two shared segments - J-K and K-L, since W3 also shares the resources of segment J-K.

In addition, this demonstrates that a single working path may be a member of a number of SMPGs. Also a single SMPG may include more than two working paths.

3.2. Shared Start and End Nodes

For the sake of the discussion of the SMP operation we designate the two end- points of the shared protection segment as a Shared Start Node (SSN) and Shared End Node (SEN). To simplify the discussion this designation is based on referencing the protection path as a pair of unidirectional LSPs.

A 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. SSN may act as a Maintenance Entity Group Intermediate Point (MIP) for each protection path sharing the same protection resources.

Similarly, a 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). A SEN acts as a MIP on each protection path that shares the protection resource.

Both end-points are involved in coordinating the use of the unidirectional shared protection segment during the shared mesh

protection operation.

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

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 3 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.

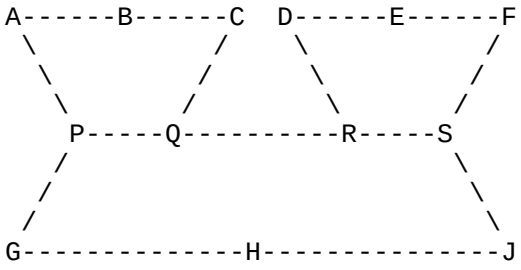


Figure 3: A More Complex Mesh Protection Example

Table 1: SSN/SEN in Figure 3

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. Connecting the end-points

The MPLS-TP Framework [RFC5921] defines the concept of a Sub-Path Maintenance Entity (SPME) and together with [RFC5586] define the use of the Generic Associated Channel (G-ACh) for communication of

MPLS-TP control protocols between the end-points of a maintenance entity, While the usual utility of a SPME is to allow tunneling of transport traffic while monitoring the segment with in-band connectivity verification messages, it is possible to use concept of a SPME to describe a LSP that is dedicated to carry a control protocol over the G-ACh between the end-points of the shared protection segment and the end-points of the protection paths within the SPMG.

For example, referring to the network in Figure 3, we would configure the following SPME (without identifying the intermediate nodes): A-P, G-P, P-Q, Q-C, D-R, G-R, S-F, S-J, R-S, and Q-J. These SPME are bidirectional LSP that are not used to carry any data traffic, only the control traffic described in Section 4.

The connection between the end-points of the shared protection segment between themselves and the end-points of the protection paths within the SPMG is to coordinate the allocation of the shared segment to a single protection path during a protection switching condition. This process is described more fully in Section 3.6

3.4. Network planning for SMP

Shared mesh protection will typically be dependent upon careful network planning. This includes:

- Preparing the working and protection paths for the different services that require protection.
- Determining which working paths are disjoint and so will not be subject to common failures. It should be clear that working paths within the same SRLG should not be included in the same SMPG.
- Identifying which protection paths share network resources and can constitute a shared protection group. Signaling or configuring the proper path information for the shared segment end-points to allow for communication between the corresponding end points of the shared segment and the protection path.
- Assigning Protection Switching Priority and a path identifier for each working path within a shared protection group.
- 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 be unprotected.
- Enabling the necessary shared mesh protection functions at the end-points of the shared protection segments. This includes

preparing the different SPME used for communication between the corresponding end points of the shared segments and the protection paths, as well as between the end-points of the shared protection segment.

Note that some control plane features of GMPLS may be used to dynamically configure 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 configured.

3.5. Preemption and race conditions

In the normal operation of SMP, when a working path triggers a protection switch, and requests allocation of the shared resources, the process should verify that the resources are available and allocate them to the requesting protection path. There are some cases where the determination of the availability is not simply determined.

Within the SMP protection domain there is a need to define a "Protection Switching Priority" for each working path. This Protection Switching Priority will be used to determine the use of the shared protection resources in cases of possible preemption. When the shared resources are in use protecting the traffic of a failed working path and a second working path fails, the SMP process should compare the Protection Switching Priority of the two working paths and if the priority of the second path is higher than the priority of the currently protected traffic, then this second path will preempt the currently protected traffic. If the second path has a lower or equal priority to the currently protected traffic, then the second path is locked-out of the protection resources.

The Protection Switching Priority may be provisioned by the network management system or configured by some other mechanism that is outside the scope of this document.

There is an additional case where the SMP process needs to make a determination of which working path should be allocated the shared resources. This is the case of multiple working paths triggering a protection switch virtually simultaneously. This may result in a race condition where the two end-points of the shared protection segment ostensibly receive requests from two different working paths. By default, working paths with equal priority results in first-come first-served recovery. If multiple working paths request protection switching simultaneously, a pre-defined identifier assigned to each working path in the SMP domain MUST be used to determine the priority among them. The definition of the identifier is for further study.

3.6. SMP Protection Switching Overview

When a protection switching trigger is activated on any of the working paths within a shared protection group, then the local linear protection mechanism (in 1:1 protection mode) should cause a protection switch. If, as a result of the protection switch action, there is a need to transmit working data on the protection path then the protection path endpoint should inform the endpoint of the shared segment of the allocation of the shared resources.

At this point the shared segment endpoints should notify all of the other protection paths in the shared protection group that the resources have been allocated, which could affect the linear protection actions relative to future triggers.

3.6.1. LP Protocol extensions for shared protection

The shared mesh protection mechanism is designed to fully utilize the existing end-to-end LP switching on the working paths. These LP domains SHALL operate in revertive mode. The LP protocol should use the normal procedures for LP without any changes except support for the following additional functionalities:

- Function to generate a protection switching event message to the SEN when a switching trigger occurs at the end-to-end linear protection domain.

- Function to take a protection locking message from the SEN, and incorporate it as the Lockout of Protection (LoP) command.

- Function to notify the SEN when the shared allocated resources may be released, when the LP domain is reverting to normal state.

3.6.2. Protection switching event

If the end point of a working path detects a switching trigger, it triggers the protection switching and exchanges LP switching protocol messages with far end-point. This operation is independent of the SMP switching mechanism specified in this document.

At the same time, for the operation of SMP, the protection path end-point notifies its protection switching event to SENs 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. The event message SHALL be transmitted over the SPME, that is configured between the protection

path end-point and the SEN, using the G-ACh. When bidirectional protection switching is being used by the working path, both end nodes will transmit the event messages to their corresponding SENs using the properly configured SPME. When unidirectional protection switching and a unidirectional failure is detected, only the detecting end-point will send the messages to its corresponding SENs.

The end-point of the protection path that is becoming active (or released) sends the messages directly to each SEN. This requires that N messages are sent, where N is the number of SMPG that the working path is a member of. This, of course, implies that the end-points are pre-configured with knowledge of all SENs associated with the SPMG.

3.6.3. Protection Locking

When a SEN receives the protection switching event notifying that protection switching to the protection path has begun in an end-to-end LP domain and that the shared resources are to be allocated, it compares the Protection Switching Priority of the working path notifying the event with those of other LP domains in the same SMPG.

The SEN determines which of the LP domains (within the SPMG) have a lower or equal priority to that of the notifying LP domain. The SEN then sends a notification to the end-points of these protection paths that is equivalent to a "Lockout of Protection" operator command. This notification should prevent any protection switching actions in those LP domains. For those LP domains having higher priorities no notification is transmitted and those LP domains may continue to perform protection switching actions.

When a protection path end point receives the protection locking message from an SEN, it SHOULD react as if a LoP command was received, according to the actions dictated by the LP protocol. Since the LoP command has the highest priority in the LP switching protocol, it will inhibit any further protection switching in the LP domain.

If the LP domain that received the protection locking message is currently transmitting traffic on the protection path, it SHALL immediately stop transmitting the traffic on the protection path and release the allocated resources.

When a SEN receives a protection switching event message indicating that the shared protection resources are being released, i.e. the LP domain is reverting to normal state, it sends a protection locking message to the end points of all the protection paths in the SMPG that were previously locked (i.e. those with equal or lower priority)

to clear the LoP command. The end-point of the protection path that receives this message SHALL react as if a Clear command was received.

3.6.4. Messages between the SEN and SSN

As was pointed out in Section 3.5 there are some cases, in particular in unidirectional protection switching triggers, of simultaneous protection switching that could cause race conditions. In these use-cases there is a need for the two end nodes of the shared protection segment, i.e. the SEN and the SSN, to coordinate the selection of the LP domain that will be allocated the shared protection resources.

For this purpose, additional messages are defined that are transmitted on the SPME that is defined between the end nodes of the shared protection segment. When a SEN receives a protection switching event notification from a LP domain indicating that protection switching to the protection path has begun, it SHALL send a message to the SSN that the resources have been allocated, with an indication of the working path identifier. This allocation needs to be confirmed for cases where both end nodes report allocation to different working path identifiers.

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:

- Protection switching event message [sent from protection path to SEN]
- Protection locking message [sent from SEN to protection path]
- Protection release message [sent from SEN to protection path]
- Resource allocation(working-path identifier) [sent from SEN to SSN]
- Resource allocation acknowledge [sent from SSN to SEN]

The channel type in ACH is used to indicate that the message is a SMP protocol message. The protocol message MUST follow the ACH.

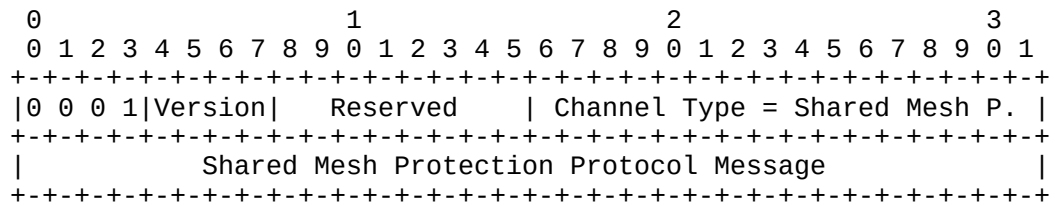


Figure 4: Shared mesh protection protocol message header

Each protocol message includes the following fields:

- o Version number
- o Identifier of the working path/LP domain - this is either the identifier of the LP domain that is sending the message or the working path that was allocated the resources (dependent upon the message)
- o Request/State field - identifies the message type as one of the messages listed above (i.e. Protection Switching Event, Protection Locking, Resource Allocation, Resource Allocation Ack)
- o Sub-request field - identifies the sub-function of the message (for example if protection path is being switched to or released for the Protection Switching Event message)

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 based on the example illustrated in Figure 3 and the following assumptions:

- The SMP domain consists of the following end-to-end LP 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)
- The SMP domain includes the following SMPG:
  - S1: LPD1 & LPD2
  - S2: LPD3 & LPD2
- Protection Switching Priority is LPD1 > LPD2 > LPD3 (i.e. LPD1 has the highest priority.)
- All working paths are protected by 1:1 bidirectional protection switching.

If a unidirectional failure occurs on W2 in the direction from node H to node G as shown in Figure 5, SMP will perform the following:

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

b. At the same time, node G transmits the protection switching event message notifying the SENs of the shared protection segments for S1 & S2, i.e. P and R, that a protection switching event occured to node.

c. SEN P compares the protection switching priority of LPD2 with those of other members of S1, i.e. LPD1. In this example, since the priority of LPD1 is higher than LPD2, SEN P does not send any message to node A.

d. SEN R compares the protection switching priority of LPD2 with those of other members of S2, i.e. 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.

e. Node D takes the protection locking message as input to the LP switching, and follows the LP procedure to process the end-to-end LoP command.

f. Since LPD2 operates in 1:1 bidirectional protection switching mode, node J performs the switching operations (i.e. switches its

bridge and selector state) to synchronize with node G, and also transmits the protection switching event message to node S and Q, which are SENs for G->H->J. Using a parallel procedure to that described in steps c & d SEN S sends the protection locking message to node F while the SEN Q does not take an action to node C.

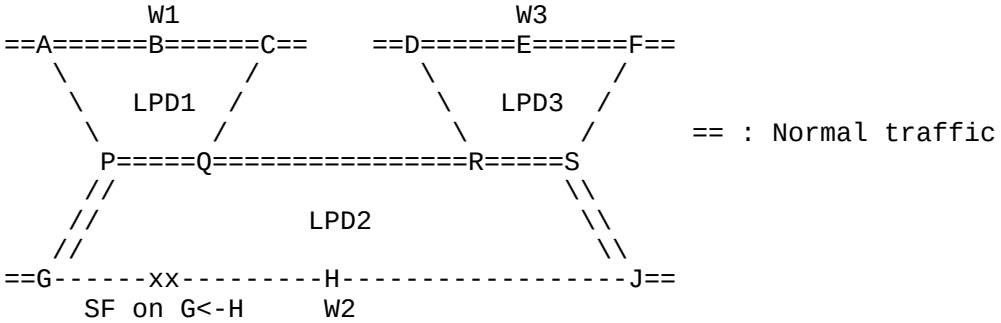


Figure 5: Shared Mesh Protection Example 1

Figure 6 shows a progression from Figure 5. While LPD2 is in protecting state with its traffic transported on protection path P2, another unidirectional failure occurs on 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 transmits the protection switching event message notifying SEN for S1, i.e. node P, that a protection switching event occurred.

c. SEN P compares the protection switching priority of LPD1 with those of the other members in S1, in this case LPD2. In this example, since the priority of LPD2 is lower than LPD1, SEN P sends the protection locking message requesting LoP to node G.

d. Node G accepts the protection locking message as input to linear protection switching, and follows LP procedure to process the 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 may be used for this, but this discussion is out of scope for this document.

e. As node G changes its bridge and selector states from protection to working, it will transmit the protection switching event message to the SENs of S1 & S2, i.e. P & R, notifying that the shared protection resources should be released.

f. SEN P compares the protection switching priority of LPD2 with the other members of S1, i.e. LPD1, and does not transmit any message to node A, but SEN R sends the protection locking message requesting clearance of LoP to node D, after comparing the protection switching priorities of the members of S2.

g. Node D accepts the message as input to the linear protection switching, and follows the LP procedures to clear the LoP command.

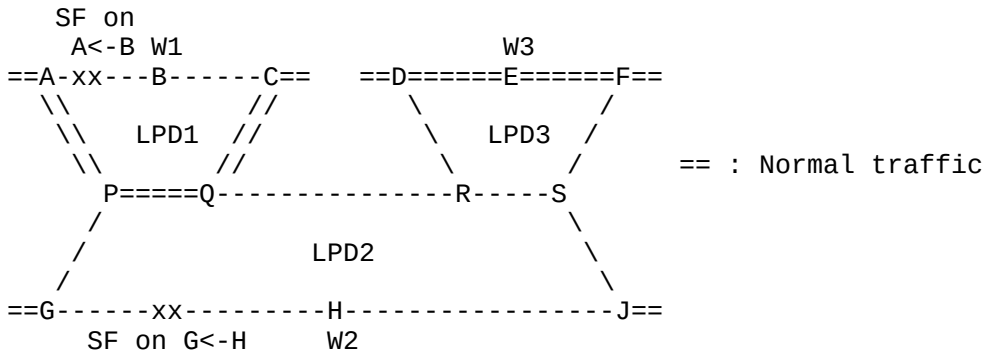


Figure 6: Shared Mesh Protection Example 2

6. Manageability Considerations

To be added in future version.

7. IANA Considerations

To be added in future version.

8. Security 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", RFC 2119, BCP 14, March 1997.

[RFC5654] Niven-Jenkins, B., Nadeau, T., and C. Pignataro, "Requirements for the Transport Profile of MPLS", RFC 5654, April 2009.

9.2. Informative References

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

[RFC3985] Bryant, S. and P. Pate, "Pseudo Wire Emulation Edge-to-Edge (PWE3) Architecture", RFC 3985, March 2005.

[RFC5921] Bocci, M., Bryant, S., Frost, D., and L. Levrau, "MPLS-TP Framework", RFC 5921, July 2010.

[RFC6372] Sprecher, N. and A. Farrel, "MPLS-TP Survivability Framework", RFC 6372, Sept 2011.

[RFC5085] Nadeau, T. and C. Pignataro, "Pseudo Wire (PW) Virtual Circuit Connectivity Verification ((VCCV): A Control Channel for Pseudowires", RFC 5085, Dec 2007.

[RFC5586] Bocci, M., Vigoureux, M., and S. Bryant, "MPLS Generic Associated Channel", RFC 5586, June 2009.

[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.

[G.808.1] SG15, "Generic Protection Switching - Linear trail and subnetwork protection", ITU-T G.808.1, Feb 2010.

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

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

Phone: +972-9-775 1827
Email: yaacov.weingarten@nsn.com

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

Email: nurit.sprecher@nsn.com

Daniel King
Old Dog Consulting
United Kingdom

Email: daniel@olddog.co.uk

Network Working Group
Internet Draft
Intended status: Informational
Expires: April 30, 2012

L. Fang, Ed.
Cisco Systems
N. Bitar
Verizon
R. Zhang
Alcatel Lucent
M. DAIKOKU
KDDI
P. Pan
Infinera

October 31, 2011

MPLS-TP Use Cases Studies and Design Considerations
draft-fang-mpls-tp-use-cases-and-design-04.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 for the new generation of packet transport. Many service providers (SPs) are working to replace the legacy transport technologies, e.g. SONET/SDH, TDM, and ATM technologies, with MPLS-TP for packet transport, in order to achieve higher efficiency, lower operational cost, while maintaining transport characteristics.

The use cases for MPLS-TP include Metro Ethernet access and aggregation, Mobile backhaul, and packet optical transport. The design considerations discussed in this documents ranging from operational experience; standards compliance; technology maturity; end-to-end forwarding and OAM consistency; compatibility with IP/MPLS networks; multi-vendor interoperability; and optimization vs. simplicity design trade off discussion. The general design principle is to provide reliable, manageable, and scalable transport solutions.

Status of this Memo

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

MPLS-TP Use Case and Design Considerations
Expires April 2012

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

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

This Internet-Draft will expire on April 30, 2012.

Copyright Notice

Table of Contents

Requirements Language

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

1. Introduction

1.1. Background and Motivation

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

In recent years, the urgency for moving from traditional transport technologies, such as SONET/SDH, TDM/ATM, to new packet technologies

has been rising. This is largely due to the tremendous success of data services, such as IPTV and IP Video for content downloading, streaming, and sharing; rapid growth of mobile services, especially smart phone applications; the continued growth of business VPNs and residential broadband. The end of live for many legacy TDM devices and the continuing network convergence effort are also key contributing factors for transport moving toward packet

technologies. After several years of heated debate on which packet technology to use, 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 LSP, tLDP for PW, as dynamic control plane options; 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 Metro access and aggregation, Mobile Backhaul, 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 enable end-to-end packet technology 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, what type of applications need to be supported, the design objectives, the cost constrain, and the network evolution plans.

1.2. Co-authors and contributors

Luyuan Fang, Cisco Systems
Nabil Bitar, Verizon
Raymond Zhang, Alcatel Lucent
Masahiro DAIKOKU, KDDI
Ping Pan, Infinera
Mach(Guoyi) Chen, Huawei Technologies
Dan Frost, Cisco Systems
Kam Lee Yap, XO Communications
Henry Yu, Time W Telecom
Jian Ping Zhang, China Telecom, Shanghai
Nurit Sprecher, Nokia Siemens Networks
Lei Wang, Telenor

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
SM-PW	Multi-Segment PW
SS-PW	Signle-Segment PW
TDM	Time Division Multiplexing
TE	Traffic Engineering
tLDP	target LDP
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 supported.

3.3. Control Plane

MPLS-TP allowed two control plane options:

Static: Using NMS for static provisioning;
Dynamic control plane for LSP: using GMPLS, OSPF-TE, RSVP-TE for full automation;
Dymanic control plane for PW: using tLDP.
ACH concept in PW is extended to G-ACh for MPLS-TP LSP to support in-band OAM.

Both Static and dynamic control plane options must allow control plane, data plane, management 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: created AIS
- Remote Defect Indication (RDI): Extended BFD
- Lock reporting: Created Lock Instruct
- Link defect Indication: Created 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

MPLS-TP OAM tool set overview can be found at [OAM Tool Set].

3.5. Survivability

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

MPLS transport Profile Survivability Framework [RFC 6372] provides more details on the subject.

4. MPLS-TP Use Case Studies

4.1. Metro Access and Aggregation

The most common deployment cases observed in the field upto today is using MPLS-TP for Metro access and aggregation. Some SPs are building green field access and aggregation infrastructure, while others are upgrading/replacing the existing transport infrastructure with new packet technologies such as MPLS-TP.
The access and aggregation networks today can be based on ATM, TDM, MSTP, or Ethernet technologies as later development.

Some other SPs announced their plans for replacing their ATM or TDM aggregation networks with MPLS-TP technologies, simply because their 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. Operators have to move forward with the next generation packet technology, the adoption of MPLS-TP in access and aggregation becomes a natural choice. 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.

4.2. Packet Optical Transport

Many SP’s transport networks consist of both packet and optical portions. The transport operators are typically sensitive to network deployment cost and operation simplicity. MPLS-TP is therefore a natural fit in some of the transport networks, where the operators can utilize the MPLS-TP LSP’s (including the ones statically provisioned) to manage user traffic as "circuits" in both packet and optical networks.
Among other attributes, bandwidth management, protection/recovery and OAM are critical in Packet/Optical transport networks. In the context of MPLS-TP, each LSP is expected to be associated with a fixed amount of bandwidth in terms of bps and/or time-slots. OAM is to be performed on each individual LSP. For some of performance monitoring (PM) functions, the OAM mechanisms need to be able transmit and process OAM packets at very high frequency, as low as several msec’s.

Protection is another important element in transport networks. Typically, ring and linear protection can be readily applied in metro networks. However, as long-haul networks are sensitive to bandwidth cost and tend to have mesh-like topology, shared mesh protection is becoming increasing important.

Packet optical deployment plans in some SPs cases are using MPLS-TP from long haul optical packet transport all the way to the aggregation and access.

4.3. Mobile Backhaul

Wireless communication 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.

4.3.1. 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.

4.3.2. 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.

4.3.3. WiMAX Backhaul

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

5. Network Design Considerations

5.1. IP/MPLS vs. MPLS-TP

Questions one might 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. MPLS-TP is developed to meet the needs of traditional transport moving towards packet. It is designed 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 common 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.

Today, majority of the core set of MPLS-TP protocol definitions are published as IETF RFCs already. It is important to deploy the solutions based on the standards definitions, in order to ensure the compatibility between MPLS-TP and IP/MPLS networks, and the interoperability among different equipment by different vendors.

Note that using non-standards, e.g. experimental code point is not recommended practice, it bares the risk of code-point collision, as indicated by [RFC 3692]: It can lead to interoperability problems when the chosen value collides with a different usage, as it someday surely will.

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. PW Design considerations in MPLS-TP networks

In general, PW works the same as in IP/MPLS network, both SS-PW and MS-PW are supported.

For dynamic control plane, tLDP is used. For static provisioning is used, PW status is a new PW OAM feature for failure notification.

In addition, both directions of a PW must be bound to the same transport bidirectional LSP.

When multi-tier rings involved in the network topology, should S-PE be used or not? It is a design trade-off.

- Pros for using S-PE
  - Domain isolation, may facilitate trouble shooting
  - the PW failure recovery may be quicker
- Cons for using S-PE
  - Adds more complexity
  - If the operation simplicity is the high priority, some SPs choose not to use S-PE, simply forming longer path across primary and secondary rings.

Should PW protection for the same end points be considered? It is another design trade-off.

- Pros for using PW protection
  - PW is protected when both working and protect LSPs carrying the working PW fails as long as the protection PW is following a diverse LSP path from the one carrying the working PW.

- Cons for using PW protection
  - Adds more complexity, some may choose not to use if protection against single point of failure is sufficient.

5.5. Proactive and event driven MPLS-TP OAM tools

MPLS-TP provide both proactive tools and event drive OAM Tools.

E.g. in the proactive fashion, the BFD hellos can be sent every 3.3 ms as its lowest interval, 3 missed hellos would be trigger the failure protection switch over. BFD sessions should be configured for both working and protecting LSPs.

When Unidirectional Failure occurs, RDI will send the failure notification to the opposite direction to trigger both end switch over.

In the reactive fashion, when there is a fiber cut for example, LDI message would be generated from the failure point and propagate to MEP to trigger immediate switch over from working to protect path. And AIS would propagate from MIP to MEP for alarm suppression.

Should both proactive and event driven OAM tools be used? The answer is yes.

Should BFD timers be set as low as possible? It depends on the applications. In many cases, it is not necessary. The lower the times are, the faster the detection time, and also the higher resource utilization. It is good to choose a balance point.

5.6. MPLS-TP and IP/MPLS Interworking considerations

Since IP/MPLS is largely deployment in most networks, MPLS-TP and IP/MPLS interworking is a reality.

Typically, there is peer model and overlay model.

The inter-connection can be simply VLAN, or PW, or could be MPLS-TE. A separate document is addressing the in the interworking issues, please refer to the descriptions in [Interworking].

5.7. 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.

5.7.1. 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.

5.8. More on MPLS-TP Deployment Considerations

5.8.1. 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.

5.8.2. Provisioning Modes Select ion

As stated in MPLS-TP requirements [RFC 5654], 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).

6. Security Considerations

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

7. IANA Considerations

This document contains no new IANA considerations.

8. 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.

9. Informative References

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

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

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

[RFC 6372] Sprecher, N., Ferrel, A., MPLS transport Profile Survivability Framework [RFC 6372], September 2011.

[OAM Tool Set] Sprecher, N., Fang, L., "An Overview of the OAM Tool Set for MPLS Based Transport Networks, ", draft-ietf-mpls-to-oam-analysis-06.txt, Oct. 2011, work in progress.

[Interworking] Martinotti, R., et al., "Interworking between MPLS-TP and IP/MPLS", draft-martinotti-mpls-tp-interworking-02.txt, June 2011.

10. Author's Addresses

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

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

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

Henry Yu
TW Telecom
10475 Park Meadow Dr.
Littleton, CO 80124
Email: henry.yu@twtelecom.com

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

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

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

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

MPLS
Internet-Draft
Intended status: Standards Track
Expires: April 23, 2012

D. Frost, Ed.
S. Bryant, Ed.
Cisco Systems
M. Bocci, Ed.
Alcatel-Lucent
October 21, 2011

MPLS Generic Associated Channel (G-ACh) Advertisement Protocol
draft-fbb-mpls-gach-adv-00

Abstract

The MPLS Generic Associated Channel (G-ACh) provides an auxiliary logical data channel associated with a Label Switched Path (LSP), a pseudowire, or a section (link) over which a variety of protocols may flow. These protocols are commonly used to provide Operations, Administration, and Maintenance (OAM) mechanisms associated with the primary data channel. This document specifies simple procedures by which an endpoint of an LSP, pseudowire, or section may inform the other endpoints of its capabilities and configuration parameters. This information may then be used by the receiver to validate or adjust its local configuration, and by the network operator for diagnostic purposes.

Status of this Memo

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

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

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

This Internet-Draft will expire on April 23, 2012.

Table of Contents

1. Introduction

The MPLS Generic Associated Channel (G-ACh) is defined and described in [RFC5586]. It provides an auxiliary logical data channel associated with an MPLS Label Switched Path (LSP), a pseudowire, or a section (link) over which a variety of protocols may flow. A primary purpose of the G-ACh and the protocols it supports is to provide Operations, Administration, and Maintenance (OAM) capabilities associated with the underlying LSP, pseudowire, or section. Examples of such capabilities include Pseudowire Virtual Circuit Connectivity Verification (VCCV) [RFC5085], Bidirectional Forwarding Detection (BFD) for MPLS [RFC5884], and MPLS packet loss, delay, and throughput measurement [RFC6374], as well as OAM functions developed for the MPLS Transport Profile (MPLS-TP) [RFC5921].

This document specifies procedures for an MPLS Label Switching Router (LSR) to advertise its capabilities and configuration parameters, or other application-specific information, to its peers over LSPs, pseudowires, and sections. Receivers can then make use of this information to validate or adjust their own configurations, and network operators can make use of it to diagnose faults and configuration inconsistencies between endpoints.

The main principle guiding the design of the MPLS G-ACh advertisement protocol (GAP) is simplicity. The protocol provides a one-way method of distributing information about the sender. How this information is used by a given receiver is a local matter. The data elements distributed by the GAP are application-specific and, except for those associated with the GAP itself, are outside the scope of this document. An IANA registry is created to allow GAP data elements to be defined as needed.

1.1. Motivation

It is frequently useful in a network for a node to have general information about its adjacent nodes, i.e., those nodes to which it has links. At a minimum this allows a human operator or management application with access to the node to determine which adjacent nodes this node can see, which is helpful when troubleshooting connectivity problems. A typical example of an "adjacency awareness protocol" is the Link Layer Discovery Protocol [LLDP], which can provide various pieces of information about adjacent nodes in Ethernet networks, such as system name, basic functional capabilities, link speed/duplex settings, and maximum supported frame size. Such data is useful both for human diagnostics and for automated detection of configuration inconsistencies.

In MPLS networks, the G-ACh provides a convenient link-layer-agnostic

means for communication between LSRs that are adjacent at the link layer. The G-ACh advertisement protocol presented in this document thus allows LSRs to exchange information of a similar sort to that supported by LLDP for Ethernet links.

An important special case arises in networks based on the MPLS Transport Profile (MPLS-TP) [RFC5921] that do not also support IP: without IP, protocols for determining the Ethernet address of an adjacent MPLS node, such as the Address Resolution Protocol [RFC0826] and IP version 6 Neighbor Discovery [RFC4861], are not available. The G-ACh advertisement protocol can be used to discover the Ethernet MAC addresses of MPLS nodes lacking IP capability [I-D.fbb-mpls-tp-ethernet-addressing].

The applicability of the G-ACh advertisement protocol is not limited to link-layer adjacency, either in terms of message distribution or message content. The G-ACh exists for any MPLS LSP or pseudowire, so GAP messages can be exchanged with remote LSP or pseudowire endpoints. The content of GAP messages is extensible in a simple manner, and can include any kind of information that might be useful to MPLS LSRs connected by links, LSPs, or pseudowires. For example, in networks that rely on the G-ACh for OAM functions, GAP messages might be used to inform adjacent LSRs of a node's OAM capabilities and configuration parameters.

1.2. Terminology

Term	Definition
G-ACh	Generic Associated Channel
GAL	G-ACh Label
GAP	G-ACh Advertisement Protocol
LSP	Label Switched Path
LSR	Label Switching Router
OAM	Operations, Administration, and Maintenance

1.3. Requirements Language

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

2. Overview

[Editor's note: The current text allows for the fragmentation of large GAP messages, and includes header fields and procedures to support this. An alternative, however, is to simplify the protocol

by removing this capability and placing the burden of fragmentation on GAP applications. The intention of the editors is to make this simplification in the next version of the draft unless there is significant interest from the community in fragmentation support.]

The G-ACh Advertisement Protocol has a simple one-way mode of operation: a device configured to send information for a particular data channel (MPLS LSP, pseudowire, or section) transmits GAP messages over the G-ACh associated with the data channel. Each message consists of one or more fragments; each message fragment contains basic information about itself and its relation to the complete message. The payload of a GAP message is a collection of Type-Length-Value (TLV) objects, organized on a per-application basis. An IANA registry is created to identify specific applications.

Although one GAP message can contain data for several applications, the receiver maintains the data associated with each application separately. This enables the sender to transmit a targeted update that refreshes the data for a subset of applications without affecting the data of other applications.

For example, a GAP message might be sent containing the following data:

Application A: A-TLV1, A-TLV2, A-TLV3

Application B: B-TLV1

Application C: C-TLV1, C-TLV2

A second message might then be sent containing:

Application B: B-TLV1, B-TLV2

Upon receiving the second message, the receiver flushes the old data for Application B and replaces it with the new data. The data associated with Applications A and C from the first message is retained. In other words, the GAP update granularity is per-application, not per-message or per-TLV-object.

The rate at which GAP messages are transmitted is at the discretion of the sender, and may fluctuate over time as well as differ per-application. Each message contains, for each application it describes, a lifetime that informs the receiver how long to wait before discarding the data for this application.

3. Message Format

An Associated Channel Header (ACH) Channel Type has been allocated for the GAP as follows:

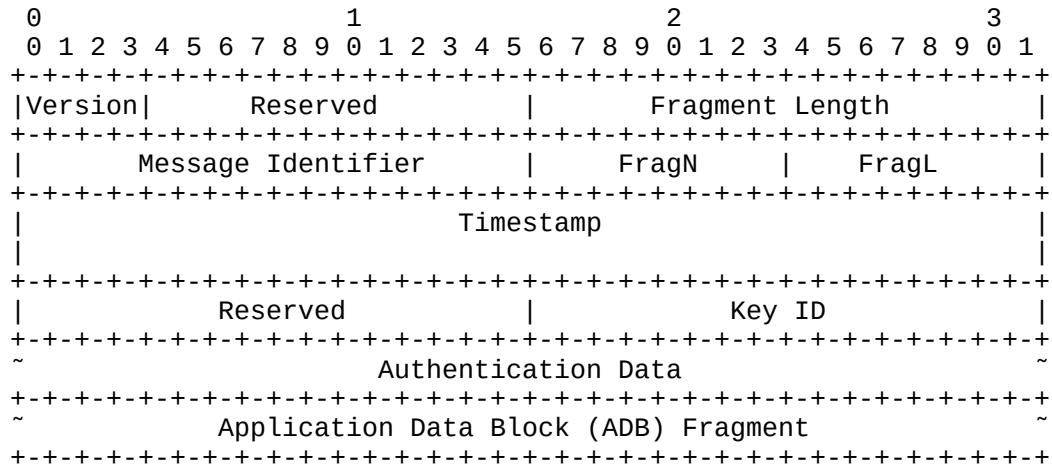
Protocol	Channel Type
G-ACh Advertisement Protocol	0xXXXX

For this Channel Type, the ACH SHALL NOT be followed by the ACH TLV Header defined in [RFC5586].

Fields in this section shown as Reserved or Resv are reserved for future specification and MUST be set to zero. All integer values for fields defined in this document SHALL be encoded in network byte order.

The payload of a GAP message is an Application Data Block (ADB) containing TLV objects for one or more applications. Since an ADB may be large, it can be broken into several fragments, with each fragment included in a separate GAP message fragment. All of these message fragments together make up a single GAP message.

The following figure shows the format of a G-ACh Advertisement Protocol message fragment, which follows the Associated Channel Header (ACH):



GAP Message Fragment Format

The meanings of the fields are:

Version: Protocol version, currently set to 0

Fragment Length: Length of this message fragment in octets

Message Identifier: Unique identifier of this message

FragN: Number of this message fragment within the total message, starting from 0

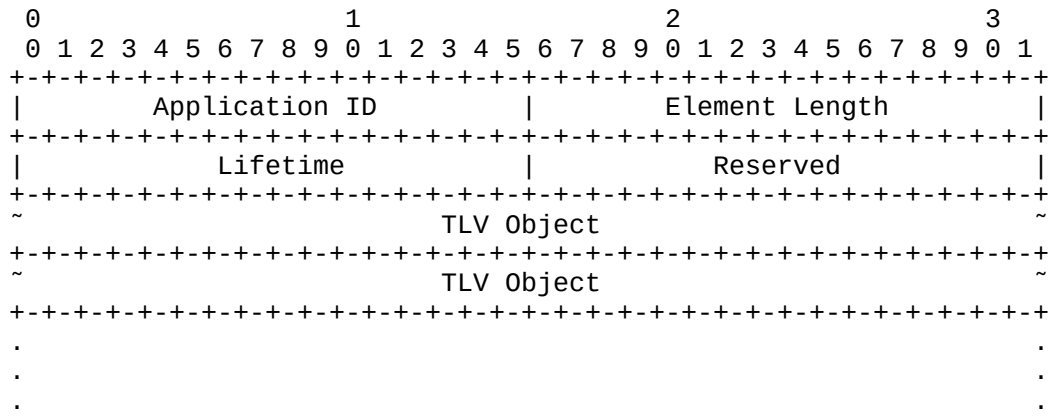
FragL: Number of the last fragment in the total message

Timestamp: 64-bit Network Time Protocol (NTP) transmit timestamp, as specified in Section 6 of [RFC5905]

Key ID: See Section 6

Authentication Data: See Section 6

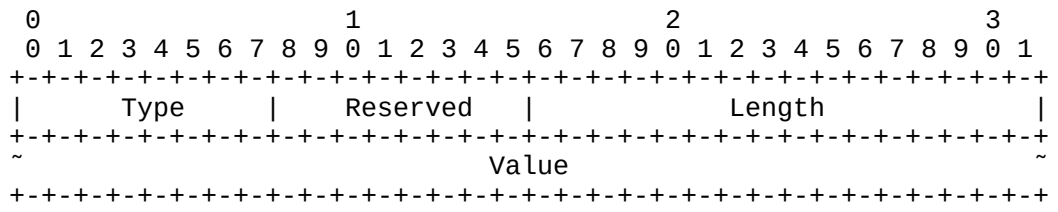
An ADB consists of one or more elements of the following format:



Application Data Block Element

In this format, the Application ID identifies the application this element describes; an IANA registry has been created to track the values for this field. Any two ADB elements in the same ADB SHALL have distinct Application IDs. The Element Length field specifies the total length in octets of this block element. The Lifetime field specifies how long, in seconds, the receiver should retain the data in this message.

The remainder of the Application Data Block element consists of a sequence of TLV objects, which are of the form:



TLV Object Format

The Type field identifies the TLV Object; an IANA registry has been created to track the values for this field, which are defined on a per-application basis. The Length field specifies the length in octets of the Value field.

It is permissible for the sequence of TLV objects in an ADB element to be empty. This is useful in conjunction with setting the Lifetime to zero in order to instruct the receiver to flush all data associated with this application.

GAP messages do not contain a checksum. If validation of message integrity is desired, the authentication procedures in Section 6 should be used.

4. G-ACh Advertisement Protocol TLVs

The GAP supports several TLV objects related to its own operation via the Application ID 0x0000. When an ADB element for the GAP is present in a GAP message, it MUST precede other elements.

4.1. Identifier TLVs

The following TLV objects are defined for purposes of conveying identification information associated with the transmitting device and the data channel:

- Interface Identifier TLV
- LSP Identifier TLV
- Pseudowire Path Identifier TLV

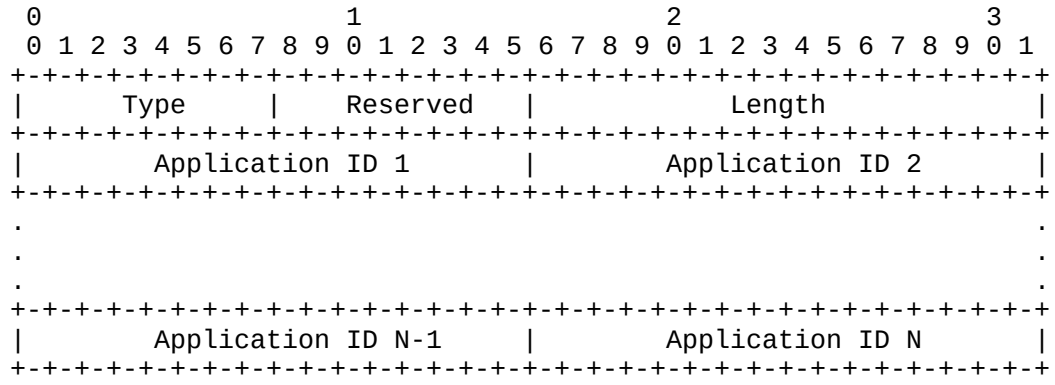
The Value portion of these identifier objects follows the format of the respective identifier as defined in [RFC6370].

The LSP and Pseudowire Path Identifiers SHOULD be present in GAP messages transmitted over LSPs and pseudowires, respectively, and

MUST NOT be present for other data channel types. The Interface Identifier SHOULD be present in GAP messages transmitted over data-link sections.

4.2. GAP Request TLV

This object is a request by the sender for the receiver to transmit an immediate unicast GAP update to the sender. If the Length field is zero, this signifies that an update for all applications is requested. Otherwise, the Value field specifies the applications for which an update is requested, in the form of a sequence of Application IDs:



GAP Request TLV Format

4.3. GAP Flush TLV

This object is an instruction to the receiver to flush the GAP data for all applications. It is a null object, i.e. its Length is set to zero. Note that data for a specific application can be flushed by sending an update for the application with the Lifetime set to zero.

The GAP Flush instruction does not apply to data contained in the message carrying the GAP Flush TLV object itself. Any application data contained in the same message SHALL be processed and retained by the receiver as usual.

4.4. GAP Suppress TLV

This object is a request to the receiver to cease sending GAP updates to the transmitter over the current channel. The object format is the same as the GAP Request TLV object. If the Length is set to zero, suppression of all GAP messages is requested; otherwise suppression of only those updates pertaining to applications listed

in the Value field is requested.

This object makes sense only for point-to-point channels or when the sender is receiving unicast GAP updates.

5. Operation

5.1. G-ACh Advertisement Message Transmission

G-ACh Advertisement Protocol message transmission SHALL operate on a per-data-channel basis and be configurable by the operator accordingly.

Because GAP message transmission may be active for many logical channels on the same physical interface, message transmission timers SHOULD be randomized across the channels supported by a given interface so as to reduce the likelihood of large synchronized message bursts.

The Message Identifier field of a GAP message MUST be the same for all message fragments of a particular message. The FragN field SHALL be set to identify the sequence number of this fragment within the total message, starting from zero. The FragL field MUST be set to the sequence number of the last fragment in this message.

The Timestamp field SHALL be set to the time at which this message fragment is transmitted.

The Key ID and Authentication Data fields SHALL be set according to the procedures in Section 6.

The Lifetime field of each Application Data Block element SHALL be set to the number of seconds the receiver is advised to retain the data associated with this message and application.

Lifetimes SHOULD be set in such a way that at least three updates will be sent prior to Lifetime expiration. For example, if updates are sent at least every 60 seconds, a Lifetime of 185 seconds may be used.

In some cases additional reliability may be desired for the delivery of a GAP message. When this is the case, the RECOMMENDED procedure is to send three instances of the message in succession, separated by a delay appropriate to the application. This procedure SHOULD be used, if at all, only for messages that are in some sense 'exceptional'; for example when sending a flush instruction following device reset.

5.2. G-ACh Advertisement Message Reception

Upon receiving a G-ACh Advertisement Protocol message containing data for a set of applications, the receiver MUST discard any earlier data retained for each application in the set, and SHOULD retain the new data associated with each application in the set by this message for the number of seconds specified by the Lifetime field, or until a newer message describing the application is received.

The receiver MAY make use of the application data contained in a GAP message to perform some level of autoconfiguration, for example if the application is an OAM protocol. The implementation SHOULD, however, take care to prevent cases of oscillation resulting from each endpoint attempting to adjust its configuration to match the other. Any such autoconfiguration based on GAP information MUST be disabled by default.

6. Message Authentication

The GAP message header provides a means of authenticating GAP message fragments and ensuring their integrity. This is accomplished by including, in the Authentication Data field, the output of a cryptographic hash function, the input to which is the message fragment together with a secret key known only to the sender and receiver. Upon receipt of the message, the receiver computes the same hash and compares the result with the hash value in the message; if the hash values are not equal, the message is discarded.

The remainder of this section gives the details of this procedure, which is based on the procedures for generic cryptographic authentication for the Intermediate System to Intermediate System (IS-IS) routing protocol as described in [RFC5310].

6.1. Authentication Key Identifiers

An Authentication Key Identifier (Key ID) is a 16-bit tag shared by the sender and receiver that identifies a set of authentication parameters. These parameters are not sent over the wire; they are assumed to be associated, on each node, with the Key ID by external means, such as via explicit operator configuration or a separate key-exchange protocol. Multiple Key IDs may be active on the sending and receiving nodes simultaneously, in which case the sender locally selects a Key ID from this set to use in an outbound message. This capability facilitates key migration in the network.

The parameters associated with a Key ID are:

- Authentication Algorithm: This signifies the authentication algorithm to use to generate or interpret authentication data. At present, the following values are possible: HMAC-SHA-1, HMAC-SHA-224, HMAC-SHA- 256, HMAC-SHA-384, and HMAC-SHA-512.

- Authentication Keystring: A secret string that forms the basis for the crytographic key used by the Authentication Algorithm.

6.2. Authentication Process

The authentication process for GAP message fragments is straightforward. First, a Key ID is associated on both the sending and receiving nodes with a set of authentication parameters. Following this, when the sender generates a GAP message fragment, it sets the Key ID field accordingly. (The length of the Authentication Data field is also known at this point, because it is a function of the Authentication Algorithm.) The sender then computes a hash for the message fragment as described below, and fills the Authentication Data field with the hash value. The message fragment is then sent.

When the message fragment is received, the receiver computes a hash for it as described below. The receiver compares its computed value to the hash value received in the Authentication Data field. If the two hash values are equal, authentication of the message fragment is considered to have succeeded; otherwise it is considered to have failed.

This process suffices to ensure the authenticity and integrity of message fragments, but is still vulnerable to a replay attack, in which a third party captures a message fragment and sends it on to the receiver at some later time. The GAP message header contains a Timestamp field which can be used to protect against replay attacks. To achieve this protection, the receiver checks that the time recorded in the timestamp field of a received and authenticated GAP message fragment corresponds to the current time, within a reasonable tolerance that allows for message propagation delay, and accepts or rejects the message fragment accordingly.

If the clocks of the sender and receiver are not synchronized with one another, then the receiver must perform the replay check against its best estimate of the current time according to the sender’s clock. The timestamps that appear in GAP messages can be used to infer the approximate clock offsets of senders and, while this does not yield high-precision clock synchronization, it suffices for purposes of the replay check with an appropriately chosen tolerance.

6.3. Hash Computation

In the algorithm description below, the following nomenclature, which is consistent with [FIPS-198], is used:

Symbol	Definition
H	The specific hash algorithm, e.g. SHA-256
K	The Authentication Keystring
Ko	The cryptographic key used with the hash algorithm
B	The block size of H, measured in octets rather than in bits. Note that B is the internal block size, not the hash size. This is equal to 64 for SHA-1 and SHA-256, and to 128 for SHA-384 and SHA-512.
L	The length of the hash, measured in octets rather than in bits
XOR	The exclusive-or operation
Opad	The hexadecimal value 0x5c repeated B times
Ipad	The hexadecimal value 0x36 repeated B times
Apad	hexadecimal value 0x878FE1F3 repeated (L/4) times

1. Preparation of the Key

   In this application, Ko is always L octets long.

   If the Authentication Keystring (K) is L octets long, then Ko is equal to K. If the Authentication Keystring (K) is more than L octets long, then Ko is set to H(K). If the Authentication Keystring (K) is less than L octets long, then Ko is set to the Authentication Keystring (K) with zeros appended to the end of the Authentication Keystring (K) such that Ko is L octets long.

2. First Hash

   First, the Authentication Data field is filled with the value Apad.

   Then, a first hash, also known as the inner hash, is computed as follows:

   First-Hash = H(Ko XOR Ipad || (GAP Message Fragment))

   Here the GAP Message Fragment is the portion of the packet that follows the Associated Channel Header.

3. Second Hash

Then a second hash, also known as the outer hash, is computed as follows:

Second-Hash = H(Ko XOR Opad || First-Hash)

4. Result

The resulting second hash becomes the authentication data that is sent in the Authentication Data field of the GAP message header. The length of the Authentication Data field is always identical to the message digest size of the specific hash function H that is being used.

This also means that the use of hash functions with larger output sizes will increase the size of the GAP message fragment as transmitted on the wire.

7. Link-Layer Considerations

When the GAP is used to support device discovery on a data link, GAP messages must be sent in such a way that they can be received by other listeners on the link without the sender first knowing the link-layer addresses of the listeners. In short, they must be multicast. Considerations for multicast MPLS encapsulation are discussed in [RFC5332]. For example, Section 8 of [RFC5332] describes how destination Ethernet MAC addresses are selected for multicast MPLS packets. Since a GAP packet transmitted over a data link contains just one label, the G-ACh Label (GAL) with label value 13, the correct destination Ethernet address for frames carrying GAP packets intended for device discovery, according to these selection procedures, is 01-00-5e-80-00-0d.

8. Security Considerations

G-ACh Advertisement Protocol messages contain information about the sending device and its configuration, which is sent in cleartext over the wire. If an unauthorized third party gains access to the MPLS data plane or the lower network layers between the sender and receiver, it can observe this information. In general, however, the information contained in GAP messages is no more sensitive than that contained in other protocol messages, such as routing updates, which are commonly sent in cleartext. No attempt is therefore made to guarantee confidentiality of GAP messages.

A more significant potential threat is the transmission of GAP messages by unauthorized sources, or the unauthorized manipulation of messages in transit; this can disrupt the information receivers hold about legitimate senders. To protect against this threat, message authentication procedures are specified in this document that enable receivers to ensure the authenticity and integrity of GAP messages. These procedures include the means to protect against replay attacks, in which a third party captures a legitimate message and "replays" it to a receiver at some later time.

9. IANA Considerations

This document requests that IANA allocate an entry in the Pseudowire Associated Channel Types registry [RFC5586] for the G-ACh Advertisement Protocol, as follows:

Value	Description	TLV Follows	Reference
(TBD)	G-ACh Advertisement Protocol	No	(this draft)

This document also requests that IANA create a new registry, "G-ACh Advertisement Protocol Applications and Data Types", with fields and initial allocations as follows:

Application ID	Description	Type Name	Type ID	Reference
0x0000	G-ACh Advertisement Protocol	GAP Request	0	(this draft)
		GAP Flush	1	(this draft)
		GAP Suppress	2	(this draft)

The allocation policy for this registry is IETF Review.

10. References

10.1. Normative References

[FIPS-198] US National Institute of Standards and Technology, "The Keyed-Hash Message Authentication Code (HMAC)", FIPS PUB 198, March 2002.

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate

Requirement Levels", BCP 14, RFC 2119, March 1997.

[RFC5332] Eckert, T., Rosen, E., Aggarwal, R., and Y. Rekhter, "MPLS Multicast Encapsulations", RFC 5332, August 2008.

[RFC5586] Bocci, M., Vigoureux, M., and S. Bryant, "MPLS Generic Associated Channel", RFC 5586, June 2009.

[RFC5905] Mills, D., Martin, J., Burbank, J., and W. Kasch, "Network Time Protocol Version 4: Protocol and Algorithms Specification", RFC 5905, June 2010.

[RFC6370] Bocci, M., Swallow, G., and E. Gray, "MPLS Transport Profile (MPLS-TP) Identifiers", RFC 6370, September 2011.

10.2. Informative References

[LLDP] IEEE, "Station and Media Access Control Connectivity Discovery (802.1AB)", September 2009.

[RFC0826] Plummer, D., "Ethernet Address Resolution Protocol: Or converting network protocol addresses to 48.bit Ethernet address for transmission on Ethernet hardware", STD 37, RFC 826, November 1982.

[RFC4861] Narten, T., Nordmark, E., Simpson, W., and H. Soliman, "Neighbor Discovery for IP version 6 (IPv6)", RFC 4861, September 2007.

[RFC5085] Nadeau, T. and C. Pignataro, "Pseudowire Virtual Circuit Connectivity Verification (VCCV): A Control Channel for Pseudowires", RFC 5085, December 2007.

[RFC5310] Bhatia, M., Manral, V., Li, T., Atkinson, R., White, R., and M. Fanto, "IS-IS Generic Cryptographic Authentication", RFC 5310, February 2009.

[RFC5884] Aggarwal, R., Kompella, K., Nadeau, T., and G. Swallow, "Bidirectional Forwarding Detection (BFD) for MPLS Label Switched Paths (LSPs)", RFC 5884, June 2010.

[RFC5921] Bocci, M., Bryant, S., Frost, D., Levrau, L., and L. Berger, "A Framework for MPLS in Transport Networks", RFC 5921, July 2010.

[RFC6374] Frost, D. and S. Bryant, "Packet Loss and Delay Measurement for MPLS Networks", RFC 6374, September 2011.

Authors' Addresses

Dan Frost (editor)
Cisco Systems

Email: danfrost@cisco.com

Stewart Bryant (editor)
Cisco Systems

Email: stbryant@cisco.com

Matthew Bocci (editor)
Alcatel-Lucent

Email: matthew.bocci@alcatel-lucent.com

MPLS
Internet-Draft
Intended status: Standards Track
Expires: April 23, 2012

D. Frost, Ed.
S. Bryant, Ed.
Cisco Systems
M. Bocci, Ed.
Alcatel-Lucent
October 21, 2011

MPLS-TP Next-Hop Ethernet Addressing
draft-fbb-mpls-tp-ethernet-addressing-00

Abstract

The Multiprotocol Label Switching (MPLS) Transport Profile (MPLS-TP) is the set of MPLS protocol functions applicable to the construction and operation of packet-switched transport networks. This document presents considerations for link-layer addressing of Ethernet frames carrying MPLS-TP packets.

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 April 23, 2012.

1. Introduction

The MPLS Transport Profile (MPLS-TP) [RFC5921] is the set of protocol functions that meet the requirements [RFC5654] for the application of MPLS to the construction and operation of packet-switched transport networks. The MPLS-TP data plane consists of those MPLS-TP functions concerned with the encapsulation and forwarding of MPLS-TP packets and is described in [RFC5960].

This document presents considerations for link-layer addressing of Ethernet frames carrying MPLS-TP packets. Since MPLS-TP packets are MPLS packets, existing procedures ([RFC3032], [RFC5332]) for the encapsulation of MPLS packets over Ethernet apply. Because IP functionality is optional in an MPLS-TP network, however, IP-based protocols for Media Access Control (MAC) address learning, such as the Address Resolution Protocol (ARP) [RFC0826] and IP version 6 Neighbor Discovery [RFC4861], may not be available. This document specifies the options for determination and selection of next-hop Ethernet MAC addressing under these circumstances.

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. Terminology

Term	Definition
ARP	Address Resolution Protocol
G-ACh	Generic Associated Channel
GAL	G-ACh Label
LSP	Label Switched Path
LSR	Label Switching Router
MAC	Media Access Control
MPLS-TP	MPLS Transport Profile
OAM	Operations, Administration, and Maintenance

Additional definitions and terminology can be found in [RFC5960] and [RFC5654].

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 [RFC2119].

2. Point-to-Point Link Addressing

When two MPLS-TP nodes are connected by a point-to-point Ethernet link, the question arises as to what destination Ethernet Media Access Control (MAC) address should be specified in Ethernet frames transmitted to the peer node over the link. The problem of determining this address does not arise in IP/MPLS networks because of the presence of the Ethernet Address Resolution Protocol (ARP) [RFC0826] or IP version 6 Neighbor Discovery protocol [RFC4861], which allow the unicast MAC address of the peer device to be learned dynamically.

If existing mechanisms are available in an MPLS-TP network to determine the destination unicast MAC addresses of peer nodes -- for example, if the network also happens to be an IP/MPLS network -- such mechanisms SHOULD be used. The remainder of this section discusses the available options when this is not the case.

Each node MAY be statically configured with the MAC address of its peer. Note however that static MAC address configuration can present an administrative burden and lead to operational problems. For example, replacement of an Ethernet interface to resolve a hardware fault when this approach is used requires that the peer node be manually reconfigured with the new MAC address. This is especially problematic if the peer is operated by another provider.

The Ethernet broadcast address FF-FF-FF-FF-FF-FF MAY be used as the destination MAC address in frames carrying MPLS-TP packets over a link that is known to be point-to-point. This may, however, lead to excessive frame distribution and processing at the Ethernet layer. Broadcast traffic may also be treated specially by some devices and this may not be desirable for MPLS-TP data frames.

The approach which SHOULD be used, in view of these considerations, is therefore to use as the destination MAC address an Ethernet multicast address reserved for MPLS-TP for use over point-to-point links. The address allocated for this purpose by the Internet Assigned Numbers Authority (IANA) is 01-00-5E-XX-XX-XX. An MPLS-TP implementation MUST process Ethernet frames received over a point-to-point link with this destination MAC address by default.

The use of broadcast or multicast addressing is applicable only when the attached Ethernet link is known to be point-to-point. If a link is not known to be point-to-point, these forms of addressing MUST NOT be used.

3. Multipoint Link Addressing

When a multipoint Ethernet link serves as a section [RFC5960] for a point-to-multipoint MPLS-TP LSP, and multicast destination MAC addressing at the Ethernet layer is used for the LSP, the addressing and encapsulation procedures specified in [RFC5332] SHALL be used.

When a multipoint Ethernet link -- that is, a link which is not known to be point-to-point -- serves as a section for a point-to-point MPLS-TP LSP, unicast destination MAC addresses MUST be used for Ethernet frames carrying packets of the LSP. According to the discussion in the previous section, this implies the use of either static MAC address configuration or a protocol that enables peer MAC address discovery.

4. MAC Address Determination via the G-ACh Advertisement Protocol

The G-ACh Advertisement Protocol (GAP) [I-D.fbb-mpls-gach-adv] provides a simple means of informing listeners on a link of the sender’s capabilities and configuration. When used for this purpose on an Ethernet link, GAP messages are multicast to the address 01-00-5e-80-00-0d. If these messages contain the unicast MAC address of the sender, then listeners can learn this address and use it in the future when transmitting frames containing MPLS-TP packets. Since the GAP does not rely on IP, this provides a means of unicast MAC discovery for MPLS-TP nodes without IP support.

This document defines a new GAP application, "Ethernet Interface Parameters", to support the advertisement of Ethernet-specific parameters associated with the sending interface. It defines several Type-Length-Value (TLV) objects for this application, as follows:

Unicast MAC Address: The Value of this object is a canonical 48-bit Ethernet unicast MAC address assigned to one of the interfaces of the sender that is connected to this data link.

Maximum Frame Size: The Value of this object is 32-bit unsigned integer encoded in network byte order that specifies the maximum frame size supported by the sending interface, in octets.

[Editor’s note: Other objects may be added here.]

5. Security Considerations

The use of broadcast or multicast Ethernet destination MAC addresses for frames carrying MPLS-TP data packets can potentially result in such frames being distributed to devices other than the intended destination node or nodes when the Ethernet link is not point-to-point. The operator SHOULD take care to ensure that MPLS-TP nodes are aware of the Ethernet link type (point-to-point or multipoint). In the case of multipoint links, the operator SHOULD either ensure that no devices are attached to the link that are not authorized to receive the frames, or take steps to mitigate the possibility of excessive frame distribution, for example by configuring the Ethernet switch to appropriately restrict the delivery of multicast frames to authorized ports.

6. IANA Considerations

6.1. Ethernet Multicast Address Allocation

IANA is requested to allocate an Ethernet Multicast Address from the Ethernet Multicast Addresses table in the ethernet-numbers registry for use by MPLS-TP LSRs over point-to-point links as described in Section 2. The entry should specify an address of the form 01-00-5E-XX-XX-XX, a Type Field of 8847/8848, and a usage "MPLS-TP point-to-point (this draft)".

6.2. G-ACh Advertisement Protocol Allocation

IANA is requested to allocate a new Application ID in the "G-ACh Advertisement Protocol Applications and Data Types" registry, along with several associated data types, as follows:

Application ID	Description	Type Name	Type ID	Reference
(TBD)	Ethernet Interface Parameters	Unicast MAC Address	0	(this draft)
		Maximum Frame Size	1	(this draft)

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.

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

[RFC5332] Eckert, T., Rosen, E., Aggarwal, R., and Y. Rekhter, "MPLS Multicast Encapsulations", RFC 5332, August 2008.

[RFC5654] Niven-Jenkins, B., Brungard, D., Betts, M., Sprecher, N., and S. Ueno, "Requirements of an MPLS Transport Profile", RFC 5654, September 2009.

[RFC5960] Frost, D., Bryant, S., and M. Bocci, "MPLS Transport Profile Data Plane Architecture", RFC 5960, August 2010.

7.2. Informative References

[RFC0826] Plummer, D., "Ethernet Address Resolution Protocol: Or converting network protocol addresses to 48.bit Ethernet address for transmission on Ethernet hardware", STD 37, RFC 826, November 1982.

[RFC4861] Narten, T., Nordmark, E., Simpson, W., and H. Soliman, "Neighbor Discovery for IP version 6 (IPv6)", RFC 4861, September 2007.

[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

Dan Frost (editor)
Cisco Systems

Email: danfrost@cisco.com

Stewart Bryant (editor)
Cisco Systems

Email: stbryant@cisco.com

Matthew Bocci (editor)
Alcatel-Lucent

Email: matthew.bocci@alcatel-lucent.com

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: April 23, 2012

X. Fu
M. Betts
Q. Wang
ZTE
D. McDysan
A. Malis
Verizon
V. Manral
Hewlett-Packard Corp.
October 21, 2011

RSVP-TE extensions for services aware MPLS
draft-fuxh-mpls-delay-loss-rsvp-te-ext-00

Abstract

With more and more enterprises using cloud based services, the distances between the user and the applications are growing. For multiple applications such as High Performance Computing and Electronic Financial markets, the response times are critical as is packet loss, while other applications require more throughput. For example, 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 and packet loss 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 April 23, 2012.

Copyright Notice

Table of Contents

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 and improve the user's experience. 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 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.

Packet Loss constraint will be taken up in a future version of the draft.

1.1. Conventions Used in This Document

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

2. 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.

Node latency for a WAN could be ignored or even an average, however that was not true for the LAN cases. Whether the node latency should be accumulated or not depends on the implementation.

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.

2.1. Signaling Extensions

2.1.1. 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.

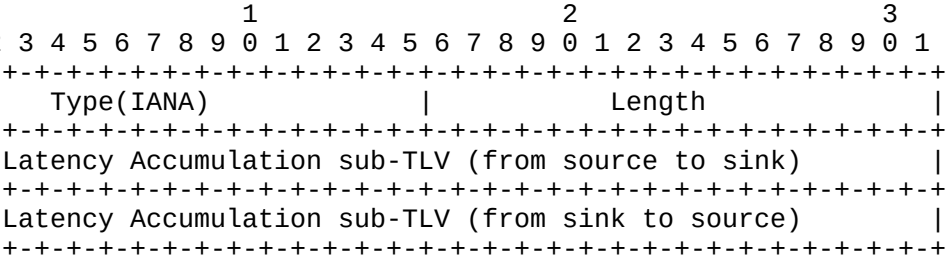


Figure 1: Format of Accumulated Latency Object

- 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.

- 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. In the case of unidirectional LSP, this sub-TLV is absent.

2.1.1.1. Latency Accumulation sub-TLV

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

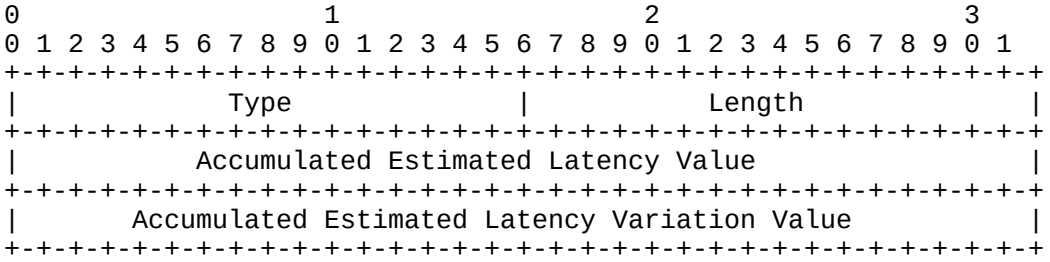


Figure 2: Format of Latency Accumulation sub-TLV

- 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.
- Length: length of the sub-TLV value in bytes.
- 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 micro-seconds and encoded as an float point value.

- Accumulated Estimated Latency Variation Value: a value indicates the sume of each links and nodes' latency variation along one direction of LSP. Since latecny variation is accumulated non-linearly. Latency variation accumulatoin should be in a lower priority. It MUST be quantified in units of nano-seconds and encoded as an float point value.

2.1.2. 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 value exceeds required 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.

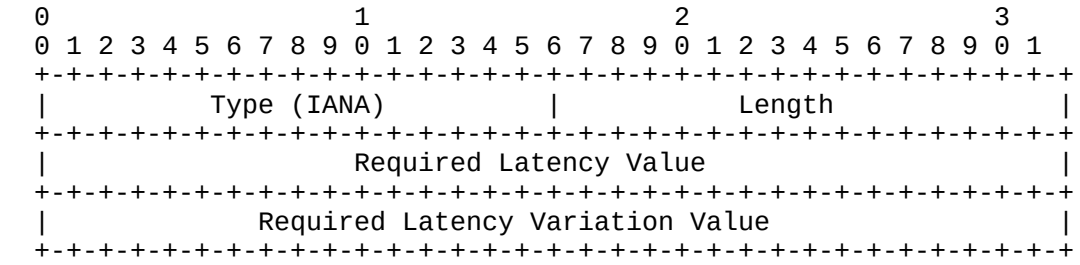


Figure 3: Required Latency Object

- Required Latency Value: The accumulated estimated latency value should not exceed this value. It MUST be quantified in units of micro-seconds and encoded as an float point value.

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

2.1.3. 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 downsteam.

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 IANA).

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 IANA).

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 IANA).

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 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 IANA).

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.

3. Performance SLA Parameters Conveying

[CL-REQ] introduces Composite Link into MPLS network. 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.

One end-to-end LSP (e.g., in IP/MPLS or MPLS-TP network) may traverse a FA-LSP of server layer (e.g., OTN rings). There will be some latency constraint requirement for server layer. RSVP-TE message also needs to carry a indication for the FA selection or FA-LSP creation.

The RSVP-TE message needs to carry a indication of request minimum

latency, maximum acceptable latency value and maximum acceptable delay variation value. The end point will take these parameters into account for selection or creation of component link, FA selection or FA-LSP.

3.1. 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.

3.1.1. 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.

- 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.

- 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.

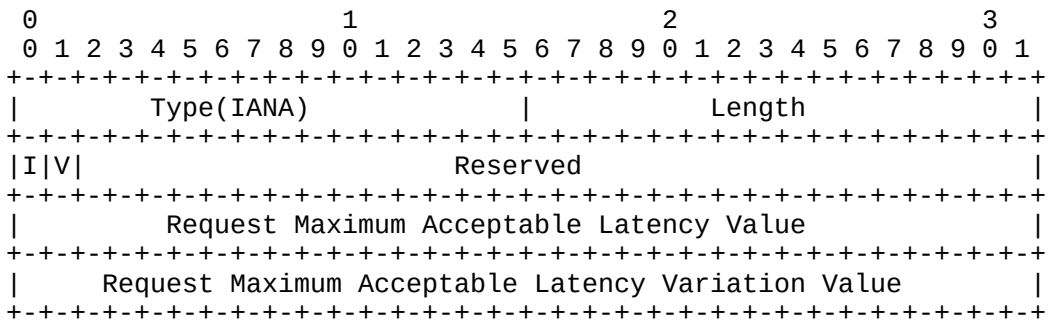


Figure 4: Format of Latency SLA Parameters TLV

- 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 micro-seconds and encoded as an float point 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 nano-seconds and encoded as an float point 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

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

Assume there are following request information.

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

Only Component link2 could be qualified.

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

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

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

Only Component link4 could be qualified.

- Request minimum latency = TRUE
- Request minimum latency variation= FALSE

- Maximum Acceptable Latency Value= 350 ms
- 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.

3.1.2. 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 IANA). 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 IANA).

4. Security Considerations

This document raises no new security issues.

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 .

[EXPRESS-PATH] S. Giacalone, "OSPF Traffic Engineering (TE) Express Path", draft-giacalone-ospf-te-express-path-01 .

[LATENCY-REQ] X. Fu, "Traffic Engineering architecture for services aware MPLS", draft-fuxh-mpls-delay-loss-te-framework-02 .

[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

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

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

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: April 10, 2012

X. Fu
ZTE
V. Manral
Hewlett-Packard Corp.
D. McDysan
A. Malis
Verizon
S. Giacalone
Thomson Reuters
M. Betts
Q. Wang
ZTE
J. Drake
Juniper Networks
October 8, 2011

Traffic Engineering architecture for services aware MPLS
draft-fuxh-mpls-delay-loss-te-framework-02

Abstract

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

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

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

Requirements Language

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

Status of this Memo

Copyright Notice

Table of Contents

1. Introduction

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

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

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

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

2. Architecture requirements overview

2.1. Communicate Latency and Loss as TE Metric

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

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

Path computation entity MUST have the capability to compute one end-to-end path with latency and packet loss constraint. For example, it has the capability to compute a route with X amount of bandwidth with less than Y ms of latency and 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.

2.2. Requirement for Composite Link

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

The solution MUST provide a means to indicate that a traffic flow should select a component link with minimum latency and/or packet loss, maximum acceptable latency and/or packet loss value and maximum acceptable delay variation value as specified by protocol. The endpoints of Composite Link will take these parameters into account for component link selection or creation. The exact details for component links will be taken up seperately and are not part of this document.

2.3. Requirement for Hierarchy LSP

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

2.4. Latency Accumulation and Verification

The solution SHOULD provide a means to accumulate (e.g., sum) of 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 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.

2.5. Restoration, Protection and Rerouting

Some customers may insist on having the ability to re-route if the latency and loss SLA is not being met. If a "provisioned" end-to-end LSP latency and/or loss could not meet the latency and loss agreement between operator and his user, the solution SHOULD support pre-defined or dynamic re-routing to handle this case based on the local policy.

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

The latency performance of pre-defined protection or dynamic re-routing LSP MUST meet the latency SLA parameter. The difference of latency value between primary and protection/restoration path SHOULD be zero.

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.

3. End-to-End Latency

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

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

Latency or one-way delay is the time it takes for a packet within a stream going from measurement point 1 to measurement point 2.

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

The total latency of an LSP consists of the sum of the latency of the

LSP hop, as well as the average latency of switching on a device, which may vary based on queuing and buffering.

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

The switching latency on a device, can be measured internally, and multiple mechanisms and data structures to do the same have been defined. Add references to papers by Verghese, Kompella, Duffield. Though the mechanisms define how to do flow based measurements, the amount of information gathered in such a case, may become too cumbersome for the Path Computation element to effectively use.

An approximation of Flow based measurement is the per DSCP value, measurement from the ingress of one port to the egress of every other port in the device.

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

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

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

4. End-to-End Jitter

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

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

There may be very less jitter on a link-hop basis.

The buffering and queuing within a device will lead to the jitter.

Just like latency measurements, jitter measurements can be appproximated as either per DSCP per port pair (Ingresss and Egress) or as per DSCP per egress port.

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

The jitter is measured in terms of 10's of nano-seconds.

5. End-to-End Loss

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

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

There may be very less loss on a link-hop basis, except in case of physical link issues.

The buffering and queuing mechanisms within a device will decide which packet is to be dropped. Just like latency and jitter measurements, the loss can best be appproximated as either per DSCP per port pair (Ingresss and Egress) or as per DSCP per egress port.

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

6. Protocol Considerations

The protocol metrics above can be sent in IGP protocol packets RFC 3630. They can then be used by the Path Computation engine to decide paths with the desired path properties.

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

A seperate measurement should be done for an LSP when it is UP. Also

LSP's path should only be recalculated when the end-to-end metrics changes in a way it becomes more than desired.

7. Control Plane Implication

7.1. Implications for Routing

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

Latency, latecny 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 latency and loss change of link MUST be notified to the IGP again. The receiving node detrimines whether the link affects any of these LSPs for which it is ingress. If there are, it must determine whether those LSPs still meet end-to-end performance objectives.

A minimum value MUST be configured to control plane. If the link performance improves beyond a configurable minimum value, it must be re-advertised. The receiving node detrimines whether a "provisioned" end-to-end LSP latency and/or loss performance is improved because of some segment performance promotion.

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

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

- o One option is that the latency and packet loss of composite link may be the range (e.g., at least minimum and maximum) latency value of all component links. It may also be the maximum latency value of all component links. In both cases, only partial information is transmited in the IGP. So the path computation entity has insufficient information to determine whether a particular path can support its latency and packet loss requirements. This leads to signaling crankback.

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

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

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

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

7.2. Implications for Signaling

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

One end-to-end LSP (e.g., in IP/MPLS or MPLS-TP network) may traverse

a FA-LSP of server layer (e.g., OTN rings). There will be some latency and packet loss constraint requirement for the segment route in server layer. So RSVP-TE message needs to carry a indication of request minimum latency and/or packet loss, maximum acceptable latency and/or packet loss value and maximum acceptable delay variation value. The boundary nodes of FA-LSP will take these parameters into account for FA selection or FA-LSP creation.

RSVP-TE needs to be extended to accumulate (e.g., sum) latency information of links and nodes along one LSP across multi-domain (e.g., Inter-AS, Inter-Area or Multi-Layer) so that 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.

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

Pre-defined protection or dynamic re-routing could be triggered to handle this case. In the case of predefined protection, large amounts of redundant capacity may have a significant negative impact on the overall network cost. Service provider may have many layers of pre-defined restoration for this transfer, but they have to duplicate restoration resources at significant cost. Solution should 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.

8. IANA Considerations

No new IANA consideration are raised by this document.

9. Security Considerations

This document raises no new security issues.

10. Acknowledgements

TBD.

11. References

11.1. Normative References

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

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

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

[RFC3477] Kompella, K. and Y. Rekhter, "Signalling Unnumbered Links in Resource ReSerVation Protocol - Traffic Engineering (RSVP-TE)", RFC 3477, January 2003.

[RFC3630] Katz, D., Kompella, K., and D. Yeung, "Traffic Engineering (TE) Extensions to OSPF Version 2", RFC 3630, September 2003.

[RFC4203] Kompella, K. and Y. Rekhter, "OSPF Extensions in Support of Generalized Multi-Protocol Label Switching (GMPLS)", RFC 4203, October 2005.

11.2. Informative References

[CL-REQ] C. Villamizar, "Requirements for MPLS Over a Composite Link", draft-ietf-rtgwg-cl-requirement-04 .

[EXPRESS-PATH]

S. Giacalone, "OSPF Traffic Engineering (TE) Express Path", draft-giacalone-ospf-te-express-path-01 .

[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

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

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

Dave McDysan
Verizon

Email: dave.mcdysan@verizon.com

Andrew Malis
Verizon

Email: andrew.g.malis@verizon.com

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

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

Malcolm Betts
ZTE

Email: malcolm.betts@zte.com.cn

Qilei Wang
ZTE

Email: wang.qilei@zte.com.cn

John Drake
Juniper Networks

Email: jdrake@juniper.net

Network Working Group
Internet-Draft
Updates: 3031 (if approved)
Intended status: Standards Track
Expires: May 3, 2012

K. Kompella
J. Drake
Juniper Networks
S. Amante
Level 3 Communications, LLC
W. Henderickx
Alcatel-Lucent
L. Yong
Huawei USA

October 31, 2011

The Use of Entropy Labels in MPLS Forwarding
draft-ietf-mpls-entropy-label-01

Abstract

Load balancing is a powerful tool for engineering traffic across a network. This memo suggests ways of improving load balancing across MPLS networks using the concept of "entropy labels". It defines the concept, describes why entropy labels are useful, enumerates properties of entropy labels that allow maximal benefit, and shows how they can be signaled and used for various applications.

Status of this Memo

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

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

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

This Internet-Draft will expire on May 3, 2012.

Copyright Notice

Table of Contents

1. Introduction

Load balancing, or multi-pathing, is an attempt to balance traffic across a network by allowing the traffic to use multiple paths. Load balancing has several benefits: it eases capacity planning; it can help absorb traffic surges by spreading them across multiple paths; it allows better resilience by offering alternate paths in the event of a link or node failure.

As providers scale their networks, they use several techniques to achieve greater bandwidth between nodes. Two widely used techniques are: Link Aggregation Group (LAG) and Equal-Cost Multi-Path (ECMP). LAG is used to bond together several physical circuits between two adjacent nodes so they appear to higher-layer protocols as a single, higher bandwidth 'virtual' pipe. ECMP is used between two nodes separated by one or more hops, to allow load balancing over several shortest paths in the network. This is typically obtained by arranging IGP metrics such that there are several equal cost paths between source-destination pairs. Both of these techniques may, and often do, co-exist in various parts of a given provider's network, depending on various choices made by the provider.

A very important requirement when load balancing is that packets belonging to a given 'flow' must be mapped to the same path, i.e., the same exact sequence of links across the network. This is to avoid jitter, latency and re-ordering issues for the flow. What constitutes a flow varies considerably. A common example of a flow is a TCP session. Other examples are an L2TP session corresponding to a given broadband user, or traffic within an ATM virtual circuit.

To meet this requirement, a node uses certain fields, termed 'keys', within a packet's header as input to a load balancing function (typically a hash function) that selects the path for all packets in a given flow. The keys chosen for the load balancing function depend on the packet type; a typical set (for IP packets) is the IP source and destination addresses, the protocol type, and (for TCP and UDP traffic) the source and destination port numbers. An overly conservative choice of fields may lead to many flows mapping to the same hash value (and consequently poorer load balancing); an overly aggressive choice may map a flow to multiple values, potentially violating the above requirement.

For MPLS networks, most of the same principles (and benefits) apply. However, finding useful keys in a packet for the purpose of load balancing can be more of a challenge. In many cases, MPLS encapsulation may require fairly deep inspection of packets to find these keys at transit LSRs.

One way to eliminate the need for this deep inspection is to have the ingress LSR of an MPLS Label Switched Path extract the appropriate keys from a given packet, input them to its load balancing function, and place the result in an additional label, termed the 'entropy label', as part of the MPLS label stack it pushes onto that packet.

The packet's MPLS entire label stack can then be used by transit LSRs to perform load balancing, as the entropy label introduces the right level of "entropy" into the label stack.

There are four key reasons why this is beneficial:

1. at the ingress LSR, MPLS encapsulation hasn't yet occurred, so deep inspection is not necessary;

2. the ingress LSR has more context and information about incoming packets than transit LSRs;

3. ingress LSRs usually operate at lower bandwidths than transit LSRs, allowing them to do more work per packet, and

4. transit LSRs do not need to perform deep packet inspection and can load balance effectively using only a packet's MPLS label stack.

This memo describes why entropy labels are needed and defines the properties of entropy labels; in particular how they are generated and received, and the expected behavior of transit LSRs. Finally, it describes in general how signaling works and what needs to be signaled, as well as specifics for the signaling of entropy labels for LDP ([RFC5036]), BGP ([RFC3107], [RFC4364]), and RSVP-TE ([RFC3209]).

1.1. Conventions used

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

The following acronyms are used:

LSR: Label Switching Router;

LER: Label Edge Router;

PE: Provider Edge router;

CE: Customer Edge device; and

FEC: Forwarding Equivalence Class.

The term ingress (or egress) LSR is used interchangeably with ingress (or egress) LER. The term application throughout the text refers to an MPLS application (such as a VPN or VPLS).

A label stack (say of three labels) is denoted by <L1, L2, L3>, where L1 is the "outermost" label and L3 the innermost (closest to the payload). Packet flows are depicted left to right, and signaling is shown right to left (unless otherwise indicated).

The term 'label' is used both for the entire 32-bit label and the 20-bit label field within a label. It should be clear from the context which is meant.

1.2. Motivation

MPLS is very successful generic forwarding substrate that transports several dozen types of protocols, most notably: IP, PWE3, VPLS and IP VPNs. Within each type of protocol, there typically exist several variants, each with a different set of load balancing keys, e.g., for IP: IPv4, IPv6, IPv6 in IPv4, etc.; for PWE3: Ethernet, ATM, Frame-Relay, etc. There are also several different types of Ethernet over PW encapsulation, ATM over PW encapsulation, etc. as well. Finally, given the popularity of MPLS, it is likely that it will continue to be extended to transport new protocols.

Currently, each transit LSR along the path of a given LSP has to try to infer the underlying protocol within an MPLS packet in order to extract appropriate keys for load balancing. Unfortunately, if the transit LSR is unable to infer the MPLS packet's protocol (as is often the case), it will typically use the topmost (or all) MPLS labels in the label stack as keys for the load balancing function. The result may be an extremely inequitable distribution of traffic across equal-cost paths exiting that LSR. This is because MPLS labels are generally fairly coarse-grained forwarding labels that typically describe a next-hop, or provide some of demultiplexing and/or forwarding function, and do not describe the packet's underlying protocol.

On the other hand, an ingress LSR (e.g., a PE router) has detailed knowledge of an packet's contents, typically through a priori configuration of the encapsulation(s) that are expected at a given PE-CE interface, (e.g., IPv4, IPv6, VPLS, etc.). They also have more flexible forwarding hardware. PE routers need this information and these capabilities to:

a) apply the required services for the CE;

b) discern the packet's CoS forwarding treatment;

c) apply filters to forward or block traffic to/from the CE;

d) to forward routing/control traffic to an onboard management processor; and,

e) load-balance the traffic on its uplinks to transit LSRs (e.g., P routers).

By knowing the expected encapsulation types, an ingress LSR router can apply a more specific set of payload parsing routines to extract the keys appropriate for a given protocol. This allows for significantly improved accuracy in determining the appropriate load balancing behavior for each protocol.

If the ingress LSR were to capture the flow information so gathered in a convenient form for downstream transit LSRs, transit LSRs could remain completely oblivious to the contents of each MPLS packet, and use only the captured flow information to perform load balancing. In particular, there will be no reason to duplicate an ingress LSR's complex packet/payload parsing functionality in a transit LSR. This will result in less complex transit LSRs, enabling them to more easily scale to higher forwarding rates, larger port density, lower power consumption, etc. The idea in this memo is to capture this flow information as a label, the so-called entropy label.

Ingress LSRs can also adapt more readily to new protocols and extract the appropriate keys to use for load balancing packets of those protocols. This means that deploying new protocols or services in edge devices requires fewer concommitant changes in the core, resulting in higher edge service velocity and at the same time more stable core networks.

2. Approaches

There are two main approaches to encoding load balancing information in the label stack. The first allocates multiple labels for a particular Forwarding Equivalance Class (FEC). These labels are equivalent in terms of forwarding semantics, but having multiple labels allows flexibility in assigning labels to flows belonging to the same FEC. This approach has the advantage that the label stack has the same depth whether or not one uses label-based load balancing; and so, consequently, there is no change to forwarding operations on transit and egress LSRs. However, it has a major

drawback in that there is a significant increase in both signaling and forwarding state.

The other approach encodes the load balancing information as an additional label in the label stack, thus increasing the depth of the label stack by one. With this approach, there is minimal change to signaling state for a FEC; also, there is no change in forwarding operations in transit LSRs, and no increase of forwarding state in any LSR. The only purpose of the additional label is to increase the entropy in the label stack, so this is called an "entropy label". This memo focuses solely on this approach.

3. Entropy Labels and Their Structure

An entropy label (as used here) is a label:

1. that is not used for forwarding;

2. that is not signaled; and

3. whose only purpose in the label stack is to provide 'entropy' to improve load balancing.

Entropy labels are generated by an ingress LSR, based entirely on load balancing information. However, they MUST NOT have values in the reserved label space (0-15). To ensure that they are not used inadvertently for forwarding, entropy labels SHOULD have a TTL of 0. The CoS field of an entropy label can be set to any value deemed appropriate.

Since entropy labels are generated by an ingress LSR, an egress LSR MUST be able to tell unambiguously that a given label is an entropy label. If any ambiguity is possible, the label above the entropy label MUST be an 'entropy label indicator' (ELI), which indicates that the following Label is an entropy label. An ELI is typically signaled by an egress LSR and is added to the MPLS label stack along with an entropy label by an ingress LSR. For many applications, the use of entropy labels is unambiguous, and an ELI is not needed. An ELI MUST have 'Bottom of Stack' (S) bit = 0 ([RFC3032]). The TTL SHOULD be set to whatever value the label above it in the stack has. The CoS field can be set to any value deemed appropriate; typically, this will be the value in the label above it in the stack.

Applications for MPLS entropy labels include pseudowires ([RFC4447]), Layer 3 VPNs ([RFC4364]), VPLS ([RFC4761], [RFC4762]) and Tunnel LSPs carrying, say, IP traffic. [I-D.ietf-pwe3-fat-pw] explains how entropy labels can be used for RFC 4447-style pseudowires, and thus

is complementary to this memo, which focuses on several other applications of entropy labels.

4. Data Plane Processing of Entropy Labels

4.1. Ingress LSR

Suppose that for a particular application (or service or FEC), an ingress LSR X is to push label stack <TL, AL>, where TL is the 'tunnel label' and AL is the 'application label'. (Note the use of the convention for label stacks described in Section 1.1. The use of a two-label stack is just for illustrative purposes.) Suppose furthermore that the egress LSR Y has told X that it is capable of processing entropy labels for this application. If X cannot insert entropy labels, it simply uses a label stack of <TL, AL> for this application. If X can insert entropy labels, it does the following for an incoming packet:

1. X identifies the application to which the packet belongs, identifies the egress LSR as Y, and thereby picks the outgoing label stack <TL, AL> to push onto the packet to send to Y.

2. X determines which keys that it will use for load balancing.

3. X, having kept state that Y can process entropy labels for this application, generates an entropy label EL (based on the output of the load balancing function).

4. If Y does not need an ELI, X pushes <TL, AL, EL> onto the packet before forwarding it to the next hop to Y.

5. If Y requires an ELI, X pushes <TL, AL, E, EL> onto the packet before forwarding it to the next hop to Y, where E is a label whose 20-bit label field is the ELI that Y signaled, and whose other fields are set as per Section 3.

Note that ingress LSR X MUST NOT include an entropy label unless the egress LSR Y for this application has indicated that it is ready to receive entropy labels. Furthermore, if Y has signaled that an ELI is needed, then X MUST include the ELI before the entropy label.

Note that the signaling and use of entropy labels in one direction (signaling from Y to X, and data path from X to Y) has no bearing on the behavior in the opposite direction (signaling from X to Y, and data path from Y to X).

4.2. Transit LSR

Transit LSRs have virtually no change in forwarding behavior. For load balancing, transit LSRs SHOULD use the whole label stack as keys for the load balancing function. Transit LSRs MUST NOT include reserved labels as input to its load balancing function. Transit LSRs MAY choose to look beyond the label stack for further keys; however, if entropy labels are being used, this may not be very useful. Looking beyond the label stack may be the simplest approach in an environment where some ingress LSRs use entropy labels and others don't, or for backward compatibility. Thus, other than using the full label stack as input to the load balancing function, transit LSRs are almost unaffected by the use of entropy labels.

4.3. Egress LSR

Suppose egress LSR Y signals that it is capable of processing entropy labels for a tunnel or an application with label L. There are three cases of interest: (a) L is the implicit NULL label, in which case an ELI is mandatory; (b) L is not the implicit NULL label and an ELI is not required (L's S bit will be used to determine whether or not there is an EL); and (c) L is not the implicit NULL label but an ELI is required.

a1) Y receives an unlabeled packet. There is obviously no EL; Y processes the packet as usual.

a2) Y receives a packet whose top label is the ELI. Y processes the TTL and CoS fields of the ELI label, ensures that the S bit is 0, then pops it, and pops the next label as well (which must be the EL), then pops it. Y processes the remaining payload as usual.

b) Y receives a packet with top label L, and an ELI is not required. Y processes L as usual; if L's S bit is 1, the label stack is done. If L's S bit is 0, the following label is the EL. Y pops the EL. Y processes the payload as usual.

c) Y receives a packet with top label L. Y processes L as usual; if L's S bit is 1, the label stack is done. If L's S bit is 0, Y checks the following label. If it is the ELI label, Y processes the TTL and CoS fields of the ELI, ensures that the S bit is 0, pops the ELI label and the following label (which is the EL), and processes the remaining payload as usual.

If there is an ELI with S bit = 1, there is an error in the label stack. Note that the TTL field of the EL (if present) will be 0; Y MUST NOT react to this.

5. Signaling for Entropy Labels

An egress LSR Y may signal to ingress LSR(s) its ability to process entropy labels on a per-application (or per-FEC) basis. As part of this signaling, Y also signals the ELI to use, if any.

In cases where an application label is used and must be the bottommost label in the label stack, Y MAY signal that no ELI is needed for that application.

In cases where no application label exists, or where the application label may not be the bottommost label in the label stack, Y MUST signal a valid ELI to be used in conjunction with the entropy label for this FEC. In this case, an ingress LSR will either not add an entropy label, or push the ELI before the entropy label. This makes the use or non-use of an entropy label by the ingress LSR unambiguous. Valid ELI label values are strictly greater than 15.

It should be noted that egress LSR Y may use the same ELI value for all applications for which an ELI is needed. The ELI MUST be a label that does not conflict with any other labels that Y has advertised to other LSRs for other applications. Furthermore, it should be noted that the ability to process entropy labels (and the corresponding ELI) may be asymmetric: an LSR X may be willing to process entropy labels, whereas LSR Y may not be willing to process entropy labels. The signaling extensions below allow for this asymmetry.

For an illustration of signaling and forwarding with entropy labels, see Figure 9.

5.1. LDP Signaling

When using LDP for signaling tunnel labels ([RFC5036]), a Label Mapping Message sub-TLV (Entropy Label sub-TLV) is used to signal an egress LSR’s ability to process entropy labels.

The presence of the Entropy Label sub-TLV in the Label Mapping Message indicates to ingress LSRs that the egress LSR can process an entropy label. In addition, the Entropy Label sub-TLV contains a label value for the ELI. If the ELI is zero, this indicates the egress doesn’t need an ELI for the signaled application; if not, the egress requires the given ELI with entropy labels. An example where an ELI is needed is when the signaled application is an LSP that can carry IP traffic.

The structure of the Entropy Label sub-TLV is shown below.

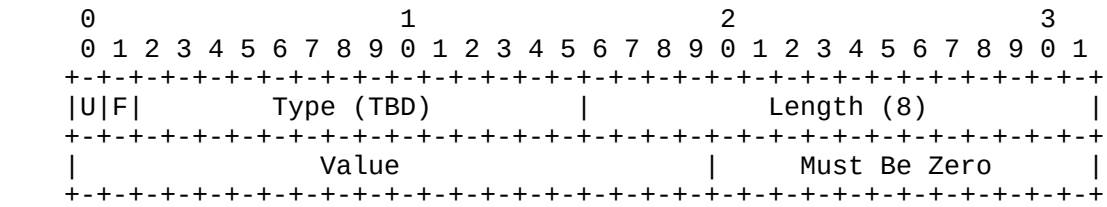


Figure 1: Entropy Label sub-TLV

where:

U: Unknown bit. This bit MUST be set to 1. If the Entropy Label sub-TLV is not understood, then the TLV is not known to the receiver and MUST be ignored.

F: Forward bit. This bit MUST be set be set to 1. Since this sub-TLV is going to be propagated hop-by-hop, the sub-TLV should be forwarded even by nodes that may not understand it.

Type: sub-TLV Type field, as specified by IANA.

Length: sub-TLV Length field. This field specifies the total length in octets of the Entropy Label sub-TLV.

Value: value of the Entropy Label Indicator Label.

5.2. BGP Signaling

When BGP [RFC4271] is used for distributing Network Layer Reachability Information (NLRI) as described in, for example, [RFC3107], [RFC4364] and [RFC4761], the BGP UPDATE message may include the Entropy Label attribute. This is an optional, transitive BGP attribute of type TBD. The inclusion of this attribute with an NLRI indicates that the advertising BGP router can process entropy labels as an egress LSR for that NLRI. If the attribute length is less than three octets, this indicates that the egress doesn’t need an ELI for the signaled application. If the attribute length is at least three octets, the first three octets encode an ELI label value as the high order 20 bits; the egress requires this ELI with entropy labels. An example where an ELI is needed is when the NLRI contains unlabeled IP prefixes.

A BGP speaker S that originates an UPDATE should only include the Entropy Label attribute if both of the following are true:

A1: S sets the BGP NEXT_HOP attribute to itself; AND

A2: S can process entropy labels for the given application.

If both A1 and A2 are true, and S needs an ELI to recognize entropy labels, then S MUST include the ELI label value as part of the Entropy Label attribute. An UPDATE SHOULD contain at most one Entropy Label attribute.

Suppose a BGP speaker T receives an UPDATE U with the Entropy Label attribute ELA. T has two choices. T can simply re-advertise U with the same ELA if either of the following is true:

B1: T does not change the NEXT_HOP attribute; OR

B2: T simply swaps labels without popping the entire label stack and processing the payload below.

An example of the use of B1 is Route Reflectors; an example of the use of B2 is illustrated in Section 9.3.1.2.

However, if T changes the NEXT_HOP attribute for U and in the data plane pops the entire label stack to process the payload, T MUST remove ELA. T MAY include a new Entropy Label attribute ELA' for UPDATE U' if both of the following are true:

C1: T sets the NEXT_HOP attribute of U' to itself; AND

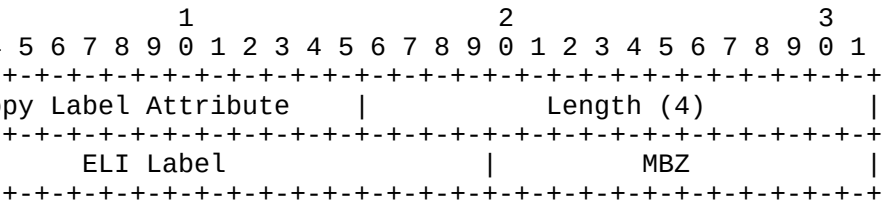
C2: T can process entropy labels for the given application.

Again, if both C1 and C2 are true, and T needs an ELI to recognize entropy labels, then T MUST include the ELI label value as part of the Entropy Label attribute.

5.3. RSVP-TE Signaling

Entropy Label support is signaled in RSVP-TE [RFC3209] using an Entropy Label Attribute TLV (Type TBD) of the LSP_ATTRIBUTES object [RFC5420]. The presence of this attribute indicates that the signaler (the egress in the downstream direction using Resv messages; the ingress in the upstream direction using Path messages) can process entropy labels. The Entropy Label Attribute contains a value for the ELI. If the ELI is zero, this indicates that the signaler doesn't need an ELI for this application; if not, then the signaler requires the given ELI with entropy labels. An example where an ELI is needed is when the signaled LSP can carry IP traffic.

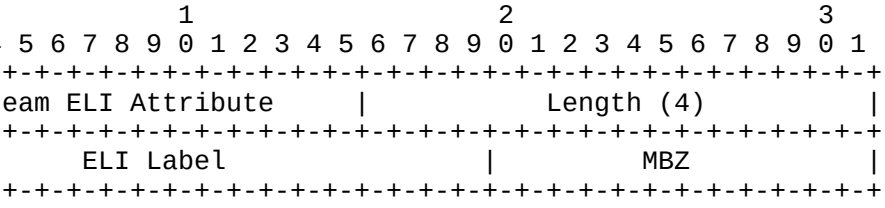
The format of the Entropy Label Attribute is as follows:



An egress LSR includes the Entropy Label Attribute in a Resv message to indicate that it can process entropy labels in the downstream direction of the signaled LSP.

An ingress LSR includes the Entropy Label Attribute in a Path message for a bi-directional LSP to indicate that it can process entropy labels in the upstream direction of the signaled LSP. If the signaled LSP is not bidirectional, the Entropy Label Attribute SHOULD NOT be included in the Path message, and egress LSR(s) SHOULD ignore the attribute, if any.

As described in Section 8, there is also the need to distribute an ELI from the ingress (upstream label allocation). In the case of RSVP-TE, this is accomplished using the Upstream ELI Attribute TLV of the LSP_ATTRIBUTES object, as shown below:



6. Operations, Administration, and Maintenance (OAM) and Entropy Labels

Generally OAM comprises a set of functions operating in the data plane to allow a network operator to monitor its network infrastructure and to implement mechanisms in order to enhance the general behavior and the level of performance of its network, e.g., the efficient and automatic detection, localization, diagnosis and handling of defects.

Currently defined OAM mechanisms for MPLS include LSP Ping/Traceroute [RFC4379] and Bidirectional Failure Detection (BFD) for MPLS [RFC5884]. The latter provides connectivity verification between the endpoints of an LSP, and recommends establishing a separate BFD session for every path between the endpoints.

The LSP traceroute procedures of [RFC4379] allow an ingress LSR to obtain label ranges that can be used to send packets on every path to the egress LSR. It works by having ingress LSR sequentially ask the transit LSRs along a particular path to a given egress LSR to return a label range such that the inclusion of a label in that range in a packet will cause the replying transit LSR to send that packet out the egress interface for that path. The ingress provides the label range returned by transit LSR N to transit LSR N + 1, which returns a label range which is less than or equal in span to the range provided to it. This process iterates until the penultimate transit LSR replies to the ingress LSR with a label range that is acceptable to it and to all LSRs along path preceding it for forwarding a packet along the path.

However, the LSP traceroute procedures do not specify where in the label stack the value from the label range is to be placed, whether deep packet inspection is allowed and if so, which keys and key values are to be used.

This memo updates LSP traceroute by specifying that the value from the label range is to be placed in the entropy label. Deep packet inspection is thus not necessary, although an LSR may use it, provided it do so consistently, i.e., if the label range to go to a given downstream LSR is computed with deep packet inspection, then the data path should use the same approach and the same keys.

In order to have a BFD session on a given path, a value from the label range for that path should be used as the EL value for BFD packets sent on that path.

As part of the MPLS-TP work, an in-band OAM channel is defined in [RFC5586]. Packets sent in this channel are identified with a reserved label, the Generic Associated Channel Label (GAL) placed at the bottom of the MPLS label stack. In order to use the inband OAM channel with entropy labels, this memo relaxes the restriction that the GAL must be at the bottom of the MPLS label stack. Rather, the GAL is placed in the MPLS label stack above the entropy label so that it effectively functions as an application label.

7. MPLS-TP and Entropy Labels

Since MPLS-TP does not use ECMP, entropy labels are not applicable to an MPLS-TP deployment.

8. Point-to-Multipoint LSPs and Entropy Labels

Point-to-Multipoint (P2MP) LSPs [RFC4875] typically do not use ECMP for load balancing, as the combination of replication and multipathing can lead to duplicate traffic delivery. However, P2MP LSPs can traverse Bundled Links [RFC4201] and LAGs. In both these cases, load balancing is useful, and hence entropy labels can be of some value for P2MP LSPs.

There are two potential complications with the use of entropy labels in the context of P2MP LSPs, both a consequence of the fact that the entire label stack below the P2MP label must be the same for all egress LSRs. First, all egress LSRs must be willing to receive entropy labels; if even one egress LSR is not willing, then entropy labels MUST NOT be used for this P2MP LSP. Second, if an ELI is required, all egress LSRs must agree to the same value of ELI. This can be achieved by upstream allocation of the ELI; in particular, for RSVP-TE P2MP LSPs, the ingress LSR distributes the ELI value using the Upstream ELI Attribute TLV of the LSP_ATTRIBUTES object, defined in Section 5.3.

With regard to the first issue, the ingress LSR MUST keep track of the ability of each egress LSR to process entropy labels, especially since the set of egress LSRs of a given P2MP LSP may change over time. Whenever an existing egress LSR leaves, or a new egress LSR joins the P2MP LSP, the ingress MUST re-evaluate whether or not to include entropy labels for the P2MP LSP.

In some cases, it may be feasible to deploy two P2MP LSPs, one to entropy label capable egress LSRs, and the other to the remaining egress LSRs. However, this requires more state in the network, more bandwidth, and more operational overhead (tracking EL-capable LSRs, and provisioning P2MP LSPs accordingly). Furthermore, this approach may not work for some applications (such mVPNs and VPLS) which automatically create and/or use P2MP LSPs for their multicast requirements.

9. Entropy Labels and Applications

This section describes the usage of entropy labels in various scenarios with different applications.

9.1. Tunnels

Tunnel LSPs, signaled with either LDP or RSVP-TE, typically carry other MPLS applications such as VPNs or pseudowires. This being the case, if the egress LSR of a tunnel LSP is willing to process entropy

labels, it would signal the need for an Entropy Label Indicator to distinguish between entropy labels and other application labels.

In the figures below, the following convention is used to depict information signaled between X and Y:

```
                    X ---------- ... ---------- Y
               app:   <--- [label L, ELI value]
```

This means Y signals to X label L for application app. The ELI value can be one of:

-: meaning entropy labels are NOT accepted;

0: meaning entropy labels are accepted, no ELI is needed; or

E: entropy labels are accepted, ELI label E is required.

The following illustrates a simple intra-AS tunnel LSP.

```
                X -------- A --- ... --- B -------- Y
      tunnel LSP L:   [TL,  E] <---  ...  <--- [TL0, E]

      IP pkt:         push <TL, E, EL> --------------->
```

Figure 2: Tunnel LSPs and Entropy Labels

Tunnel LSPs may cross Autonomous System (AS) boundaries, usually using BGP ([RFC3107]). In this case, the AS Border Routers (ASBRs) MAY simply propagate the egress LSR's ability to process entropy labels, or they MAY declare that entropy labels may not be used. If an ASBR (say A2 below) chooses to propagate the egress LSR Y's ability to process entropy labels, A2 MUST also propagate Y's choice of ELI.

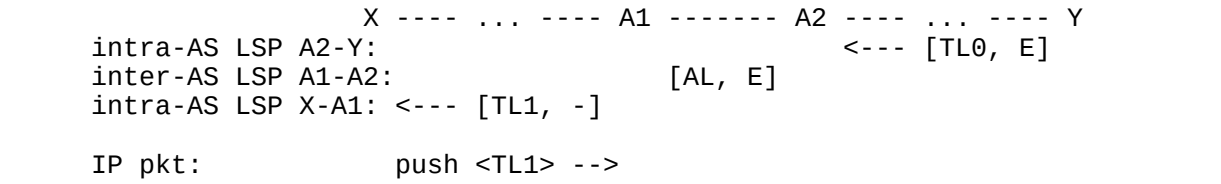
```
                X ---- ... ---- A1 ------- A2 ---- ... ---- Y
  intra-AS LSP A2-Y:                          <--- [TL0, E]
  inter-AS LSP A1-A2:              [AL, E]
  intra-AS LSP X-A1: <--- [TL1, E]

  IP pkt:          push <TL1, E, EL>
```

Here, ASBR A2 chooses to propagate Y's ability to process entropy labels, by "translating" Y's signaling of entropy label capability (say using LDP) to BGP; and A1 translate A2's BGP signaling to (say) RSVP-TE. The end-to-end tunnel (X to Y) will have entropy labels if

X chooses to insert them.

Figure 3: Inter-AS Tunnel LSP with Entropy Labels



Here, ASBR A1 decided that entropy labels are not to be used; thus, the end-to-end tunnel cannot have entropy labels, even though both X and Y may be capable of inserting and processing entropy labels.

Figure 4: Inter-AS Tunnel LSP with no Entropy Labels

9.2. LDP Pseudowires

[I-D.ietf-pwe3-fat-pw] describes the signaling and use of entropy labels in the context of RFC 4447 pseudowires, so this will not be described further here.

[RFC4762] specifies the use of LDP for signaling VPLS pseudowires. An egress VPLS PE that can process entropy labels can indicate this by adding the Entropy Label sub-TLV in the LDP message it sends to other PEs. An ELI is not required. An ingress PE must maintain state per egress PE as to whether it can process entropy labels.

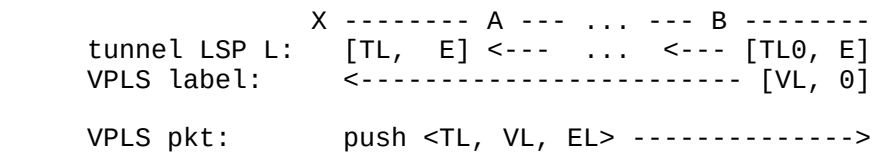


Figure 5: Entropy Labels with LDP VPLS

Note that although the underlying tunnel LSP signaling indicated the need for an ELI, VPLS packets don't need an ELI, and thus the label stack pushed by X do not have one.

[RFC4762] also describes the notion of "hierarchical VPLS" (H-VPLS). In H-VPLS, 'hub PEs' remove the label stack and process VPLS packets; thus, they must make their own decisions on the use of entropy labels, independent of other hub PEs or spoke PEs with which they exchange signaling. In the example below, spoke PEs X and Y and hub

PE B can process entropy labels, but hub PE A cannot.

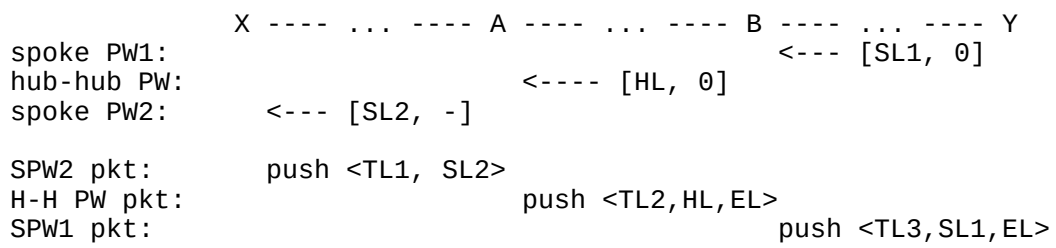


Figure 6: Entropy Labels with H-VPLS

9.3. BGP Applications

Section 9.1 described a BGP application for the creation of inter-AS tunnel LSPs. This section describes two other BGP applications, IP VPNs ([RFC4364]) and BGP VPLS ([RFC4761]). An egress PE for either of these applications indicates its ability to process entropy labels by adding the Entropy Label attribute to its BGP UPDATE message. Again, ingress PEs must maintain per-egress PE state regarding its ability to process entropy labels. In this section, both of these applications will be referred to as VPNs.

In the intra-AS case, PEs signal application labels and entropy label capability to each other, either directly, or via Route Reflectors (RRs). If RRs are used, they must not change the BGP NEXT_HOP attribute in the UPDATE messages; furthermore, they can simply pass on the Entropy Label attribute as is.

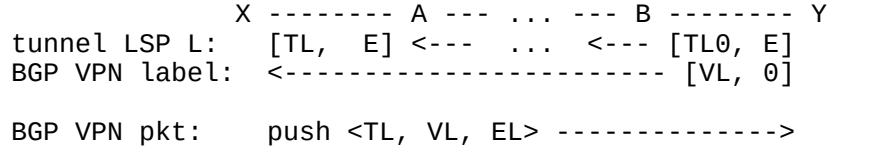


Figure 7: Entropy Labels with Intra-AS BGP apps

For BGP VPLS, the application label is at the bottom of stack, so no ELI is needed. For BGP IP VPNs, the application label is usually at the bottom of stack, so again no ELI is needed. However, in the case of Carrier's Carrier (CsC) VPNs, the BGP VPN label may not be at the bottom of stack. In this case, an ELI is necessary for CsC VPN packets with entropy labels to distinguish them from nested VPN packets. In the example below, the nested VPN signaling is not shown; the egress PE for the nested VPN (not shown) must signal

whether or not it can process egress labels, and the ingress nested VPN PE may insert an entropy label if so.

Three cases are shown: a plain BGP VPN packet, a CsC VPN packet originating from X, and a transit nested VPN packet originating from a nested VPN ingress PE (conceptually to the left of X). It is assumed that the nested VPN packet arrives at X with label stack <ZL, CVL> where ZL is the tunnel label (to be swapped with <TL, CL>) and CVL is the nested VPN label. Note that Y can use the same ELI for the tunnel LSP and the CsC VPN (and any other application that needs an ELI).

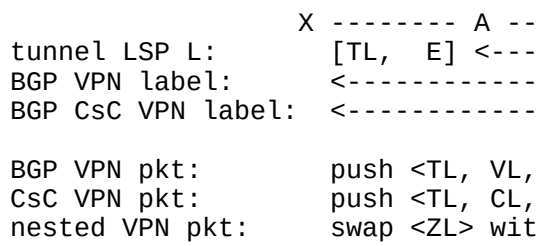


Figure 8: Entropy Labels with CoC VPN

9.3.1. Inter-AS BGP VPNs

There are three commonly used options for inter-AS IP VPNs and BGP VPLS, known informally as "Option A", "Option B" and "Option C". This section describes how entropy labels can be used in these options.

9.3.1.1. Option A Inter-AS VPNs

In option A, an ASBR pops the full label stack of a VPN packet exiting an AS, processes the payload header (IP or Ethernet), and forwards the packet natively (i.e., as IP or Ethernet, but not as MPLS) to the peer ASBR. Thus, entropy label signaling and insertion are completely local to each AS. The inter-AS paths do not use entropy labels, as they do not use a label stack.

9.3.1.2. Option B Inter-AS VPNs

The ASBRs in option B inter-AS VPNs have a choice (usually determined by configuration) of whether to just swap labels (from within the AS to the neighbor AS or vice versa), or to pop the full label stack and process the packet natively. This choice occurs at each ASBR in each direction. In the case of native packet processing at an ASBR, entropy label signaling and insertion is local to each AS and to the

inter-AS paths (which, unlike option A, do have labeled packets).

In the case of simple label swapping at an ASBR, the ASBR can propagate received entropy label signaling onward. That is, if a PE signals to its ASBR that it can process entropy labels (via an Entropy Label attribute), the ASBR can propagate that attribute to its peer ASBR; if a peer ASBR signals that it can process entropy labels, the ASBR can propagate that to all PEs within its AS). Note that this is the case even though ASBRs change the BGP NEXT_HOP attribute to "self", because of clause B2 in Section 5.2.

9.3.1.3. Option C Inter-AS VPNs

In Option C inter-AS VPNs, the ASBRs are not involved in signaling; they do not have VPN state; they simply swap labels of inter-AS tunnels. Signaling is PE to PE, usually via Route Reflectors; however, if RRs are used, the RRs do not change the BGP NEXT_HOP attribute. Thus, entropy label signaling and insertion are on a PE-pair basis, and the intermediate routers, ASBRs and RRs do not play a role.

9.4. Multiple Applications

It has been mentioned earlier that an ingress PE must keep state per egress PE with regard to its ability to process entropy labels. An ingress PE must also keep state per application, as entropy label processing must be based on the application context in which a packet is received (and of course, the corresponding entropy label signaling).

In the example below, an egress LSR Y signals a tunnel LSP L, and is prepared to receive entropy labels on L, but requires an ELI. Furthermore, Y signals two pseudowires PW1 and PW2 with labels PL1 and PL2, respectively, and indicates that it can receive entropy labels for both pseudowires without the need of an ELI; and finally, Y signals a L3 VPN with label VL, but Y does not indicate that it can receive entropy labels for the L3 VPN. Ingress LSR X chooses to send native IP packets to Y over L with entropy labels, thus X must include the given ELI (yielding a label stack of <TL, ELI, EL>). X chooses to add entropy labels on PW1 packets to Y, with a label stack of <TL, PL1, EL>, but chooses not to do so for PW2 packets. X must not send entropy labels on L3 VPN packets to Y, i.e., the label stack must be <TL, VL>.

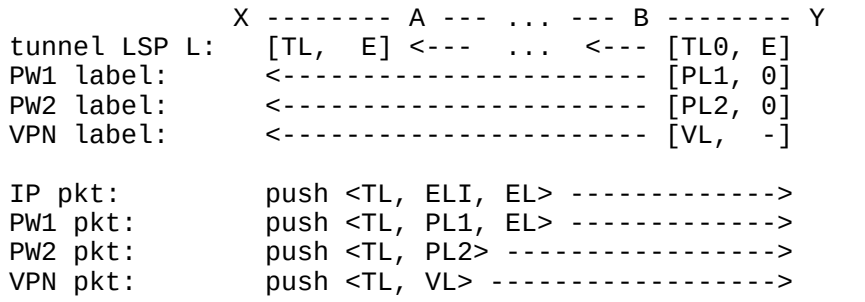


Figure 9: Entropy Labels for Multiple Applications

10. Security Considerations

This document describes advertisement of the capability to support receipt of entropy-labels and an Entropy Label Indicator that an ingress LSR may apply to MPLS packets in order to allow transit LSRs to attain better load-balancing across LAG and/or ECMP paths in the network.

This document does not introduce new security vulnerabilities to LDP. Please refer to the Security Considerations section of LDP ([RFC5036]) for security mechanisms applicable to LDP.

Given that there is no end-user control over the values used for entropy labels, there is little risk of Entropy Label forgery which could cause uneven load-balancing in the network.

If Entropy Label Capability is not signaled from an egress PE to an ingress PE, due to, for example, malicious configuration activity on the egress PE, then the PE's will fall back to not using entropy labels for load-balancing traffic over LAG or ECMP paths which, in some cases, in no worse than the behavior observed in current production networks. That said, operators are recommended to monitor changes to PE configurations and, more importantly, the fairness of load distribution over equal-cost LAG or ECMP paths. If the fairness of load distribution over a set of paths changes that could indicate a misconfiguration, bug or other non-optimal behavior on their PE's and they should take corrective action.

Given that most applications already signal an Application Label, e.g.: IPVPNs, LDP VPLS, BGP VPLS, whose Bottom of Stack bit is being re-used to signal entropy label capability, there is little to no additional risk that traffic could be misdirected into an inappropriate IPVPN VRF or VPLS VSI at the egress PE.

In the context of downstream-signaled entropy labels that require the use of an Entropy Label Indicator (ELI), there should be little to no additional risk because the egress PE is solely responsible for allocating an ELI value and ensuring that ELI label value DOES NOT conflict with other MPLS labels it has previously allocated. On the other hand, for upstream-signaled entropy labels, e.g.: RSVP-TE point-to-point or point-to-multipoint LSP's or Multicast LDP (mLDP) point-to-multipoint or multipoint-to-multipoint LSP's, there is a risk that the head-end MPLS LER may choose an ELI value that is already in use by a downstream LSR or LER. In this case, it is the responsibility of the downstream LSR or LER to ensure that it MUST NOT accept signaling for an ELI value that conflicts with MPLS label(s) that are already in use.

11. IANA Considerations

11.1. LDP Entropy Label TLV

IANA is requested to allocate the next available value from the IETF Consensus range in the LDP TLV Type Name Space Registry as the "Entropy Label TLV".

11.2. BGP Entropy Label Attribute

IANA is requested to allocate the next available Path Attribute Type Code from the "BGP Path Attributes" registry as the "BGP Entropy Label Attribute".

11.3. Attribute Flags for LSP_Attributes Object

IANA is requested to allocate a new bit from the "Attribute Flags" sub-registry of the "RSVP TE Parameters" registry.

Bit No	Name	Attribute Flags Path	Attribute Flags Resv	RRO
TBD	Entropy Label LSP	Yes	Yes	No

11.4. Attributes TLV for LSP_Attributes Object

IANA is requested to allocate the next available value from the "Attributes TLV" sub-registry of the "RSVP TE Parameters" registry.

12. Acknowledgments

We wish to thank Ulrich Drafz for his contributions, as well as the entire 'hash label' team for their valuable comments and discussion.

13. References

13.1. Normative References

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

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

[RFC3107] Rekhter, Y. and E. Rosen, "Carrying Label Information in BGP-4", RFC 3107, May 2001.

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

[RFC5420] Farrel, A., Papadimitriou, D., Vasseur, JP., and A. Ayyangarps, "Encoding of Attributes for MPLS LSP Establishment Using Resource Reservation Protocol Traffic Engineering (RSVP-TE)", RFC 5420, February 2009.

13.2. Informative References

[I-D.ietf-pwe3-fat-pw] Bryant, S., Filsfils, C., Drafz, U., Kompella, V., Regan, J., and S. Amante, "Flow Aware Transport of Pseudowires over an MPLS Packet Switched Network", draft-ietf-pwe3-fat-pw-07 (work in progress), July 2011.

[RFC4201] Kompella, K., Rekhter, Y., and L. Berger, "Link Bundling in MPLS Traffic Engineering (TE)", RFC 4201, October 2005.

[RFC4271] Rekhter, Y., Li, T., and S. Hares, "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, January 2006.

[RFC4364] Rosen, E. and Y. Rekhter, "BGP/MPLS IP Virtual Private Networks (VPNs)", RFC 4364, February 2006.

[RFC4379] Kompella, K. and G. Swallow, "Detecting Multi-Protocol Label Switched (MPLS) Data Plane Failures", RFC 4379,

February 2006.

[RFC4447] Martini, L., Rosen, E., El-Aawar, N., Smith, T., and G. Heron, "Pseudowire Setup and Maintenance Using the Label Distribution Protocol (LDP)", RFC 4447, April 2006.

[RFC4761] Kompella, K. and Y. Rekhter, "Virtual Private LAN Service (VPLS) Using BGP for Auto-Discovery and Signaling", RFC 4761, January 2007.

[RFC4762] Lasserre, M. and V. Kompella, "Virtual Private LAN Service (VPLS) Using Label Distribution Protocol (LDP) Signaling", RFC 4762, January 2007.

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

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

[RFC5586] Bocci, M., Vigoureux, M., and S. Bryant, "MPLS Generic Associated Channel", RFC 5586, June 2009.

[RFC5884] Aggarwal, R., Kompella, K., Nadeau, T., and G. Swallow, "Bidirectional Forwarding Detection (BFD) for MPLS Label Switched Paths (LSPs)", RFC 5884, June 2010.

Appendix A. Applicability of LDP Entropy Label sub-TLV

In the case of unlabeled IPv4 (Internet) traffic, the Best Current Practice is for an egress LSR to propagate eBGP learned routes within a SP's Autonomous System after resetting the BGP next-hop attribute to one of its Loopback IP addresses. That Loopback IP address is injected into the Service Provider's IGP and, concurrently, a label assigned to it via LDP. Thus, when an ingress LSR is performing a forwarding lookup for a BGP destination it recursively resolves the associated next-hop to a Loopback IP address and associated LDP label of the egress LSR.

Thus, in the context of unlabeled IPv4 traffic, the LDP Entropy Label sub-TLV will typically be applied only to the FEC for the Loopback IP address of the egress LSR and the egress LSR will not announce an entropy label capability for the eBGP learned route.

Authors' Addresses

Kireeti Kompella
Juniper Networks
1194 N. Mathilda Ave.
Sunnyvale, CA 94089
US

Email: kireeti@juniper.net

John Drake
Juniper Networks
1194 N. Mathilda Ave.
Sunnyvale, CA 94089
US

Email: jdrake@juniper.net

Shane Amante
Level 3 Communications, LLC
1025 Eldorado Blvd
Broomfield, CO 80021
US

Email: shane@level3.net

Wim Henderickx
Alcatel-Lucent
Copernicuslaan 50
2018 Antwerp
Belgium

Email: wim.henderickx@alcatel-lucent.com

Lucy Yong
Huawei USA
1700 Alma Dr. Suite 500
Plano, TX 75075
US

Email: lucyyong@huawei.com

Internet Engineering Task Force
Internet-Draft
Intended status: Standards Track
Expires: May 3, 2012

Q. Zhao
Huawei Technology
L. Fang
C. Zhou
Cisco Systems
L. Li
China Mobile
N. So
Verizon Business
R. Torvi
Juniper Networks
October 31, 2011

LDP Extensions for Multi Topology Routing
draft-ietf-mpls-ldp-multi-topology-01.txt

Abstract

Multi-Topology (MT) routing is supported in IP through extension of IGP protocols, such as OSPF and IS-IS. It would be advantageous to extend Multiprotocol Label Switching (MPLS), using Label Distribution Protocol (LDP), to support multiple topologies. These LDP extensions, known as Multiple Topology Label Distribution Protocol (MT LDP), would allow the configuration of multiple topologies within an MPLS LDP enabled network.

This document describes the protocol extensions required to extend the existing MPLS LDP signalling protocol for creating and maintaining LSPs in an MT environment.

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 May 3, 2012.

Copyright Notice

Table of Contents

1. Terminology

Terminology used in this document

MT-ID: A 12 bit value to represent Multi-Topology ID.

Default Topology: A topology that is built using the MT-ID value 0.

MT topology: A topology that is built using the corresponding MT-ID.

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. 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, this solution utilises the same MT-ID of this topology in LDP. Thus 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 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 managment, 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 forwrding 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 tolology is defined as all low latency links on the network and VoIP data packets are assinged 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 prtocol (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 distingushed by the quality of service (QoS) assinged to it.

This document describes the protocol extensions required to extend the existing MPLS LDP signalling protocol for creating and maintaining LSPs in an MT environment.

3. Requirements

MPLS-MT may 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.

The following specific requirements and objectives have been defined

in order to provide the functionality described above, and facilitate service provider configuration and operation.

- o Deployment of MPLS-MT within existing MPLS networks should be possible, with MPLS-MT non-capable nodes existing with MPLS-MT capable nodes.

- o Minimise configuration and operation complexity of MPLS-MT across the network.

- o The MPLS-MT solution SHOULD NOT require data-plane modification.

- o The MPLS-MT solution MUST support multiple topologies. Allowing a an MPLS LSP to be established across a specific, or set of, multiple topologies.

- o Control and filtering of LSPs using explicitly including or excluding multiple topologies MUST be supported.

- o The MPLS-MT solution MUST be capable of supporting QoS mechanisms.

[Editors Note - We expect these base MPLS-MT protocol requirements to be evolved over the next few versions of this document. Note that all Editors notes will be deleted before publication of the document]

3.1. Application Scenarios

3.1.1. Simplified Data-plane

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

3.1.2. 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.

3.1.3. Using MT for mLDP Protection

Fro 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 priamry LSP in the appplicaitons such 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 guranteed to have the link and node diversity from the primary LSP. By using MPLS-LDP-MT, two topology can be configured with complete link and node diversity, where the primary and secondary LSP can be set up independantly within each topology. The two LSPs setup by this mechanism can protect each other end-to-end.

3.1.4. Service Separation

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

3.1.5. An Alternative inter-AS VPN Solution

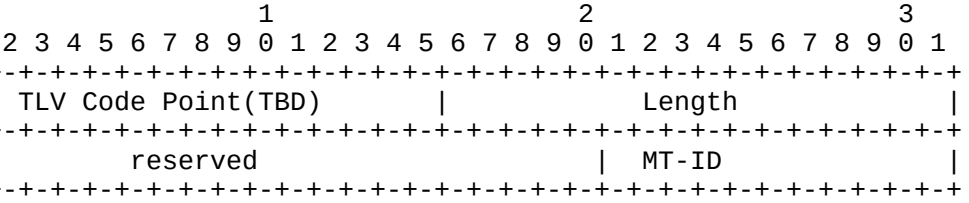
When the lsp is crossing multiple domains for the inter-as VPN scenarios, the LSP setup process can be done by configuring a set of routers which are in different domains into a new single domain with a new topology ID using the LDP multiple topology. All the routers belong this new topology will be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

3.2. Associating a FEC or group of FECs with MT-ID

This section describes multiple 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. There are also other choices such as defining new address family or associate the MPLS MT-ID with each FEC element in the FEC TLV. In this version, we discuss the first two choices, and in the future versions, we will add the discussions for other choices into the draft.

3.2.1. MT-ID TLV

The new TLV for MT-ID is defined as below:



where:

U and F bits:
As specified in [RFC5036].

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"

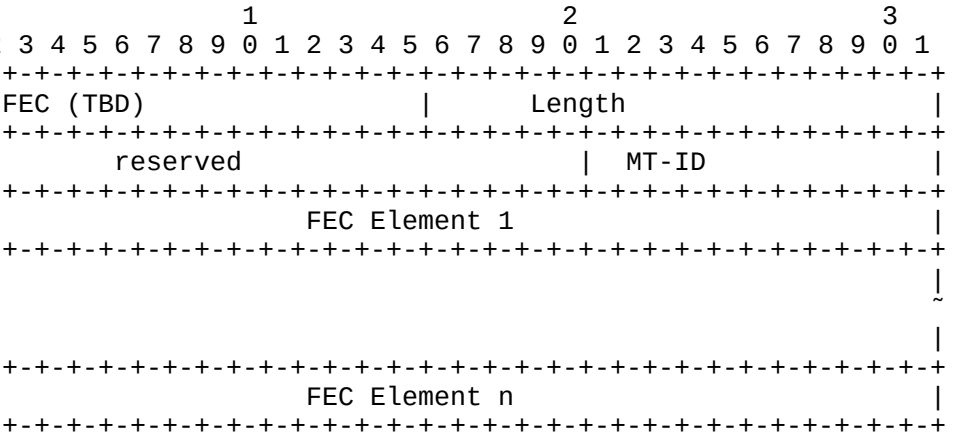
The message with inserted MT-ID TLV associates a FEC in same message to the topology identified by MT-ID.

Figure 1: MT-ID TLV Format

3.2.2. 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

3.3. LDP MT Capability Advertisement

The LDP MT capability can be advertised either during the LDP session initializatin 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).

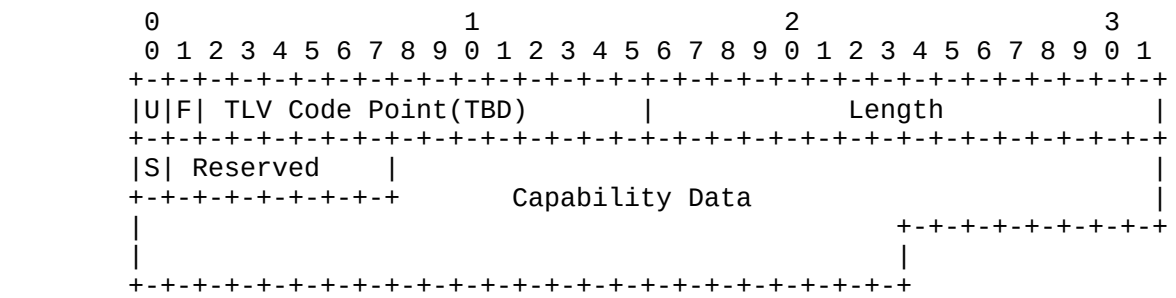
If an LSR is changed from MT capable to non-MT capable, it may initiate withdraw of all label mapping for existing LSPs of all non-default MTs. Alternatively, it may wait until the routing update to withdraw FEC and release the label mapping for existing LSPs of specific MT.

There will be case where IGP is MT capable but MPLS is not and the handling procedure for this case is TBD.

3.3.1. 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 [RFC5036] as below:



where:

U and F bits:
As specified in [RFC5036].

TLV Code Point:
The TLV type which identifies a specific capability. The "IANA Considerations" section of [RFC5036] 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

3.3.2. Post Session Setup

During the normal operating stage of LDP sessions, the capability message defined in the [RFC5036] will be used with an LDP MT capability TLV.

The format of the Capability message is as follows:

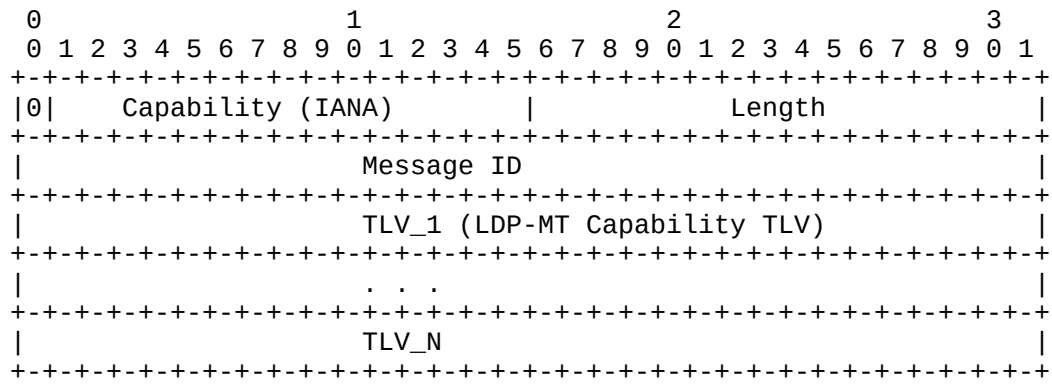


Figure 4: LDP CAP Format

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.

3.4. LDP Sessions

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

[Editors Note - This section requires further discussion]

3.5. Reserved MT ID Values

Certain MT topologies are assigned to serve pre-determined purposes:

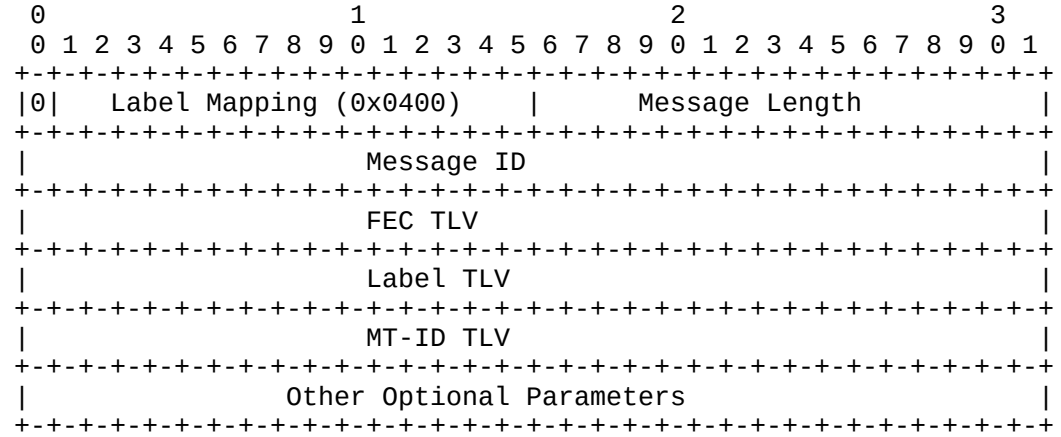
[Editors Note - This section requires further discussion]

3.6. LDP Messages with FEC TLV and MT-ID TLV

3.6.1. 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 Message ID TLV	4	See below
Hop Count TLV	1	See below
Path Vector TLV	variable	See below
MT TLV	variable	See below

MT TLV
see the definition section for this new TLV.

Figure 5: Label Mapping Message

3.6.2. 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:

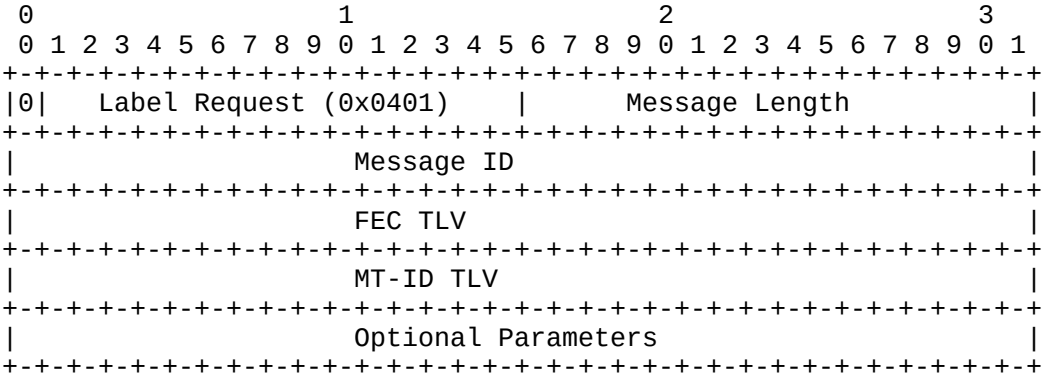


Figure 6: Label Request Message

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.

3.6.3. 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:

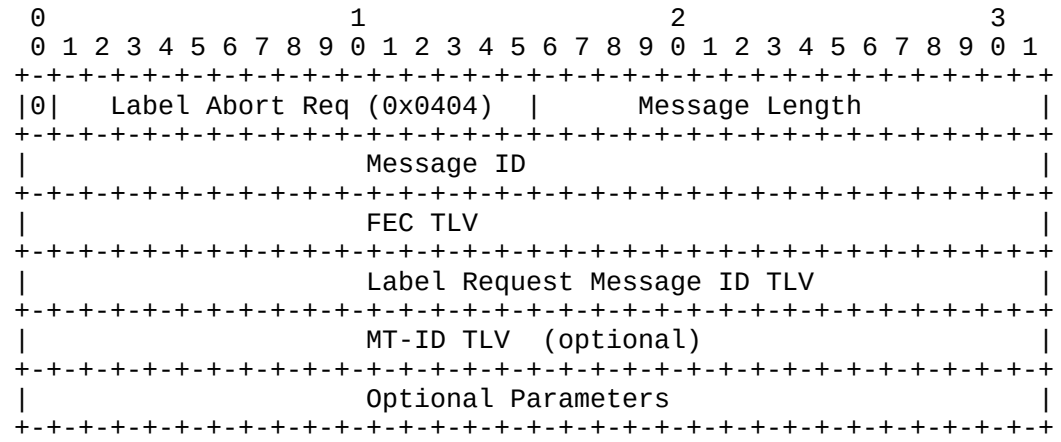


Figure 7: Label Abort Request Message

3.6.4. 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 paramters field.

The encoding for the Label Withdraw Message is:

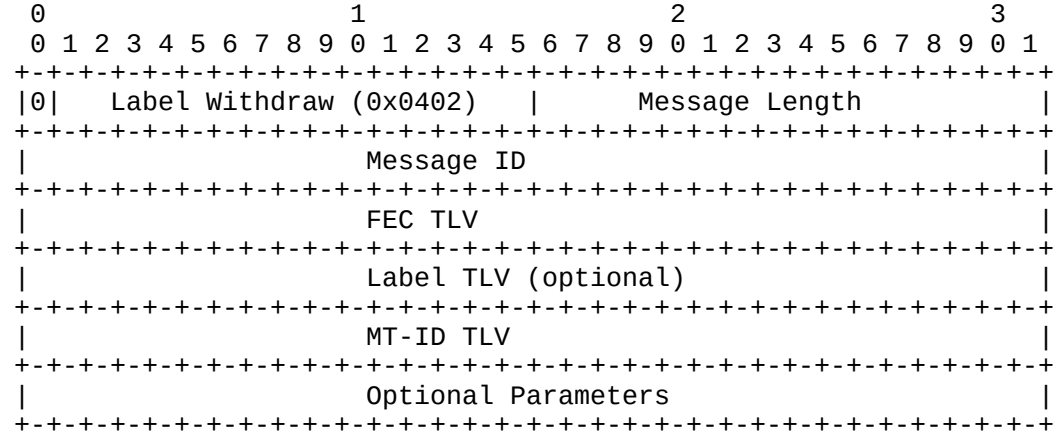


Figure 8: Label Withdraw Message

3.6.5. 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 paramers field.

The encoding for the Label Release Message is:

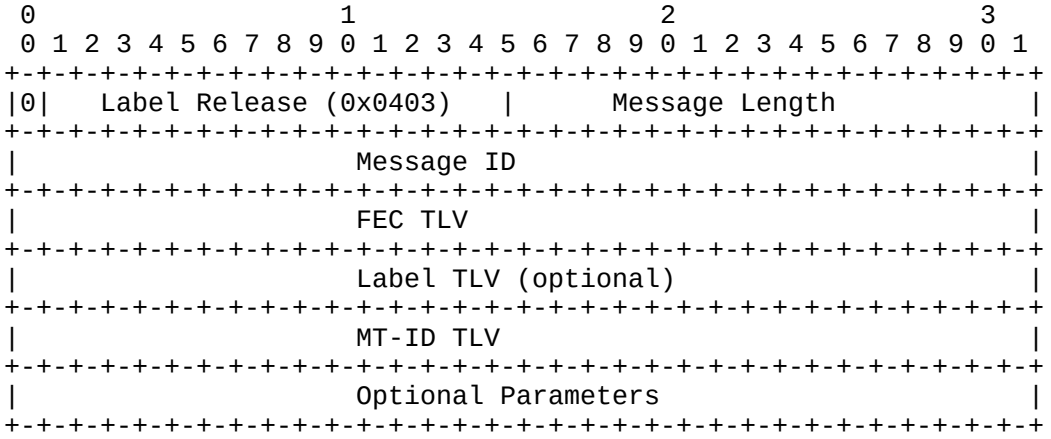


Figure 9: Label Release Message

3.7. Session Initialization Message with MT Capability

The session initializtion message is extended to contain the LDP MT capability as an optional parameter. The extended session initialization message has the format below.

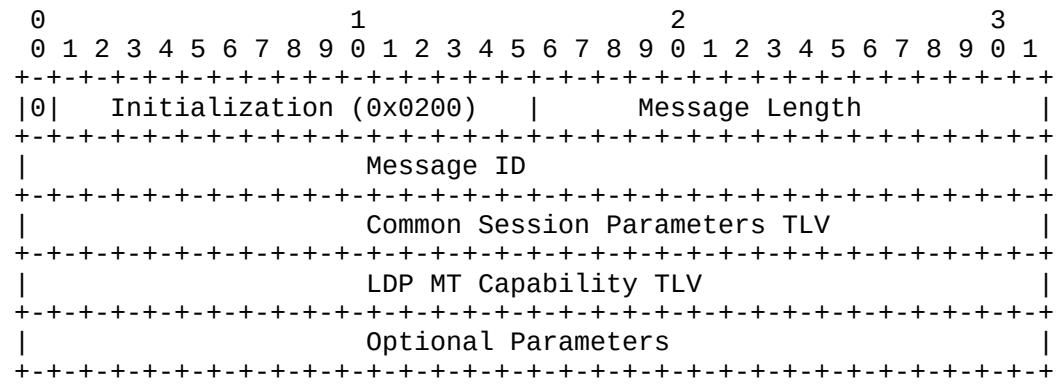


Figure 10: Session Initialization Message with MT Capability

3.8. MT Applicability on FEC-based features

3.8.1. Typed Wildcard Prefix FEC Element

RFC-5918 extends base LDP and defines Typed Wildcard FEC Element framework [RFC5918]. Typed Wildcard FEC element can be used in any LDP message to specify a wildcard operation/action for given type of FEC.

The impact of the MT extensions proposed in document on the procedures for Typed Wildcard Prefix FEC element depends on the MPLS MT-ID representation mechanism we chose at the end.

For example, if the MPLS-MT ID TLV option is the final choice, then the procedures defined in [RFC5918] apply as-is to Prefix FEC element or the Prefix FEC element along with the MPLS MT-ID TLV. For instance, upon local un-configuration of topology "x", an LSR may send wildcard label withdraw with MT-ID TLV "x" to withdraw all its labels from peer that were advertised under the scope of topology "x".

3.8.2. End-of-LIB

[RFC5919] specifies extensions and procedures for an LDP speaker to signal its convergence for given FEC type towards a peer.

The impact of the MT extensions proposed in document on the procedures for End-of-LIB depends on the MPLS MT-ID representation mechanism we chose at the end.

For example, if the MPLS-MT ID TLV option is the final choice, the

procedures defined in [RFC5919] apply as-is to Prefix FEC element or the Prefix FEC element along with the MPLS MT-ID TLV. This means that an LDP speaker MAY signal its IP convergence using Typed Wildcard Prefix FEC element, and its MT IP convergence per topology using the Typed Wildcard Prefix FEC element along with the MPLS MT-ID TLV.

3.9. MPLS Forwarding in MT

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

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

3.10. 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.

3.11. IANA Considerations

[Editors Note - This section requires further discussion]

3.12. Acknowledgement

The authors would like to thank Dan Tappan, Nabil Bitar, Huang Xin, Daniel King and Eric Rosen for their valuable comments on this draft.

4. References

4.1. 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.

[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.

4.2. Informative References

Authors' Addresses

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

Email: quintin.zhao@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
China

Email: lichenyj@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

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

Email: E-mail: skraza@cisco.com

Ning So
Verizon 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: rtorvi@juniper.net

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: May 3, 2012

M. Chen
W. Cao
Huawei Technologies Co., Ltd
S. Ning
Verizon Inc.
F. Jounay
France Telecom
S. Delord
Alcatel-Lucent
October 31, 2011

Return Path Specified LSP Ping
draft-ietf-mpls-return-path-specified-lsp-ping-04.txt

Abstract

This document defines extensions to the failure-detection protocol for Multiprotocol Label Switching (MPLS) Label Switched Paths (LSPs) known as "LSP Ping" that allow selection of the LSP to use for the echo reply return path. Enforcing a specific return path can be used to verify bidirectional connectivity and also increase LSP ping robustness. It may also be used by Bidirectional Forwarding Detection (BFD) for MPLS bootstrap signaling thereby making BFD for MPLS more robust.

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

Table of Contents

1. Introduction

This document defines extensions to the failure-detection protocol for Multiprotocol Label Switching (MPLS) Label Switched Paths (LSPs) known as "LSP Ping" [RFC4379] that can be used to specify the return paths for the echo reply message, increasing the robustness of LSP Ping, reducing the opportunity for error, and improving the reliability of the echo reply message. A new reply mode, which is referred to as "Reply via specified path", is added and a new Type-Length-Value (TLV), which is referred to as Reply Path (RP) TLV, is defined in this memo.

With the extensions described in this document, a bidirectional LSP and a pair of unidirectional LSPs (one for each direction) could both be tested with a single operational action, hence providing better control plane scalability. The defined extensions can also be utilized for creating a single Bidirectional Forwarding Detection (BFD)[RFC5880], [RFC5884]session for a bidirectional LSP or for a pair of unidirectional LSPs (one for each direction).

In this document, term bidirectional LSP includes the co-routed bidirectional LSP defined in [RFC3945] and the associated bidirectional LSP that is constructed from a pair of unidirectional LSPs (one for each direction), and which are associated with one another at the LSP’s ingress/egress points [RFC5654].

2. Problem Statements and Solution Overview

MPLS LSP Ping is defined in [RFC4379]. It can be used to detect data path failures in all MPLS LSPs, and was originally designed for unidirectional LSPs.

LSP are increasingly being deployed to provide bidirectional services. The co-routed bidirectional LSP is defined in [RFC3471] and [RFC3473], and the associated bidirectional LSP is defined in [RFC5654]. With the deployment of such services, operators have a desire to test both directions of a bidirectional LSP in a single operation.

Additionally, when testing a single direction of an LSP (either a unidirectional LSP, or a single direction of a bidirectional LSP) using LSP Ping, the validity of the result may be affected by the success of delivering the echo reply message. Failure to exchange these messages between the egress Label Switching Router (LSR) and the ingress LSR can lead to false negatives where the LSP under test is reported as "down" even though it is functioning correctly.

2.1. Limitations of Existing Mechanisms for Bidirectional LSPs

With the existing LSP Ping mechanisms as defined in [RFC4379], operators have to enable LSP detection on each of the two ends of a bidirectional LSP independently. This not only doubles the workload for the operators, but may also bring additional difficulties when checking the backward direction of the LSP under the following conditions:

1. The LSR that the operator logged on to perform the checking operations might not have out-of-band connectivity to the LSR at the far end of the LSP. That can mean it is not possible to check the return direction of a bidirectional LSP in a single operation - the operator must log on to the LSR at the other end of the LSP to test the return direction.

2. The LSP being tested might be an inter-domain/inter-AS LSP where the operator of one domain/AS may have no right to log on to the LSR at the other end of the LSP since this LSR resides in another domain/AS. That can make it completely impossible for the operator to check the return direction of a bidirectional LSP.

Associated bidirectional LSPs have the same issues as those listed for co-routed bidirectional LSPs.

This document defines a mechanism to allow the operator to request that both directions of a bidirectional LSP be tested by a single LSP Ping message exchange.

2.2. Limitations of Existing Mechanisms for Handling Unreliable Return Paths

[RFC4379] defines 4 reply modes:

1. Do not reply
2. Reply via an IPv4/IPv6 UDP packet
3. Reply via an IPv4/IPv6 UDP packet with Router Alert
4. Reply via application level control channel.

Obviously, the issue of the reliability of the return path for an echo reply message does not apply in the first of these cases.

[RFC4379] states that the third mode may be used when the IP return path is deemed unreliable. This mode of operation requires that all intermediate nodes must support the Router Alert option and must understand and know how to forward MPLS echo replies.

This is a rigorous requirement in deployed IP/MPLS networks especially since the return path may be through legacy IP-only routers. Furthermore, for inter-domain LSPs, the use of the Router Alert option may encounter significant issues at domain boundaries where the option is usually stripped from all packets. Thus, the use of this mode may itself introduce issues that lead to the echo reply messages not being delivered.

And in any case, the use modes 2 or 3 cannot guarantee the delivery of echo responses through an IP network that is fundamentally unreliable. The failure to deliver echo response messages can lead to false negatives making it appear that the LSP has failed.

Allowing the ingress LSR to control the path used for echo reply messages, and in particular forcing those messages to use an LSP rather than being sent through the IP network, enables an operator to apply an extra level of deterministic process to the LSP Ping test.

This document defines extensions to LSP Ping that can be used to specify the return paths of the echo reply message in an LSP echo request message.

3. Extensions

LSP Ping defined in [RFC4379] is carried out by sending an echo request message. It carries the Forwarding Equivalence Class (FEC) information of the tested LSP which indicates which MPLS path is being verified, along the same data path as other normal data packets belonging to the FEC.

LSP Ping [RFC4379] defines four reply modes that are used to direct the egress LSR in how to send back an echo reply. This document defines a new reply mode, the Reply Via Specified Path mode. This new mode is used to direct the egress LSR of the tested LSP to send the echo reply message back along the path specified in the echo request message.

In addition, a new TLV, the Reply Path (RP) TLV, is defined in this document. The RP TLV consists of one or more sub-TLVs that can be used to carry the specified return path information to be used by the echo reply message.

3.1. Reply Via Specified Path mode

A new reply mode is defined to be carried in the Reply Mode field of the LSP Ping echo request message.

The recommended value of the Reply Via Specified Path mode is 5 (This is to be confirmed by the IANA).

Value	Meaning
5	Reply via specified path

The Reply Via Specified Path mode is used to notify the remote LSR receiving the LSP Ping echo request message to send back the echo reply message along the specified paths carried in the Reply Path TLV.

3.2. Reply Path (RP) TLV

The Reply Path (RP) TLV is optionally included in an echo request message. It carries the specified return paths that the echo reply message is required to follow. The format of RP TLV is as follows:

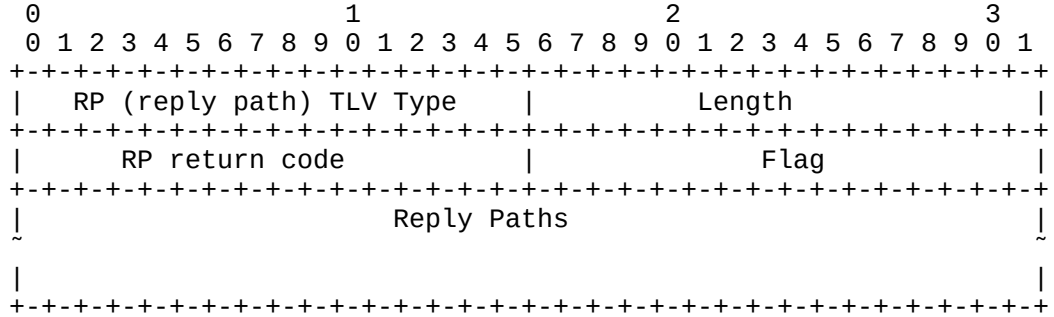


Figure3 IPv6 PSN Tunnel sub-TLV format

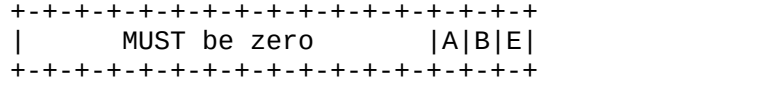
RP TLV Type field is 2 octets in length, and the type value is TBD by IANA.

The Length field is 2 octets in length. It defines the length in octets of the RP return code, Flag and Reply Paths fields.

RP return code is 2 octets in length. It is defined for the egress LSR of the forward LSP to report the results of RP TLV processing and return path selection. When sending echo request, these codes MUST be set to zero. RP return code only used when sending echo reply, and it MUST be ignored when processing echo request message. This document defines the following RP return codes:

Value	Meaning
0	No return code
1	Malformed RP TLV was received
2	One or more of the sub-TLVs in RP TLV was not understood
3	The echo reply was sent successfully using the specified RP
4	The specified RP was not found, the echo reply was sent via other LSP
5	The specified RP was not found, the echo reply was sent via IP path
6	The Reply mode in echo request was not set to 5(replay via specified path) although RP TLV exists
7	RP TLV was missing in echo request

Flag field is also 2 octets in length, it is used to notify the egress how to process the Reply Paths field when performing return path selection. The Flag field is a bit vector and has following format:



A (Alternative path): the egress LSR MUST select a non-default path as the return path. This is very useful when reverse default path problems are suspected which can be confirmed when the echo reply is forced to follow a non-default return path. If A bit is set, there is no need to carry any specific reply path sub-TLVs.

B (Bidirectional): the return path is required to follow the reverse direction of the tested bidirectional LSP.

E (Exclude): the return path is required to exclude the paths that are identified by the reply path sub-TLVs carried in the Reply Paths field. This is very useful when one or more previous LSP Ping attempts failed. By setting this E bit and carrying the previous failed reply path sub-TLVs, a new LSP Ping echo request could be used to help the egress LSR to select another candidate path when sending echo reply message.

A bit MUST NOT be set when any one of other two bits (B bit and E bit) set.

The Reply Paths field is variable in length. It has several nested sub-TLVs that describe the specified paths the echo reply message is required to follow. When the Reply Mode field is set to "Reply via specified path" in an LSP echo request message, the RP TLV MUST be present.

3.3. RP TLV sub-TLVs

Each of the FEC sub-TLVs for the Target FEC Stack TLV[RFC4379] is applicable to be a sub-TLV for inclusion in the RP TLV for expressing a specific return path.

In addition, three more new sub-TLVs are defined: IPv4 RSVP Tunnel sub-TLV, IPv6 RSVP Tunnel sub-TLV, and RP TC (Traffic Class) sub-TLV. Detailed definition is in the following sections.

With the Return Path TLV flags and the sub-TLVs defined for the Target FEC Stack TLV and in this document, it could provide following options for return paths specifying:

1. Specify a particular LSP as return path

   - use those sub-TLVs defined for the Target FEC Stack TLV

2. Specify a more generic tunnel FEC as return path

   - use the IPv4/IPv6 RSVP Tunnel sub-TLVs defined in Section 3.3.1 and Section 3.3.2 of this document

3. Specify the reverse path of the bidirectional LSP as return path

   - use B bit defined in Section 3.2 of this document.

4. Force return path to non-default path

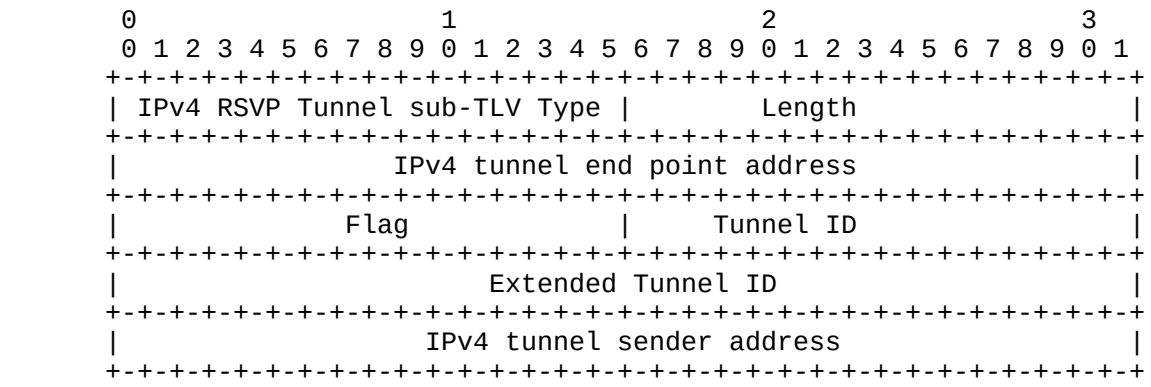
   - use A bit defined in Section 3.2 of this document.

5. Allow any LSPs except specific or general ones as return path

   - use E bit (Section 3.2 of this document) and combine with the specific paths identified by the sub-TLVs carried in Reply Path field.

3.3.1. IPv4 RSVP Tunnel sub-TLV

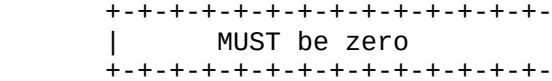
The IPv4 RSVP Tunnel sub-TLV is used in the RP TLV to allow the operator to specify a more generic tunnel FEC other than a particular LSP as the return path. The egress LSR chooses any LSP from the LSPs that have the same Tunnel attributes and satisfy the conditions carried in the Flag field. The format of IPv4 RSVP Tunnel sub-TLV is as follows:



The IPv4 RSVP Tunnel sub-TLV is derived from the RSVP IPv4 FEC TLV that is defined in Section 3.2.3 [RFC4379]. All fields have the same semantics as defined in [RFC4379] except that the LSP-ID field is omitted and a new Flag field is defined.

The IPv4 RSVP Tunnel sub-TLV Type field is 2 octets in length, and the recommended type value is 18 (to be confirmed by IANA).

The Flag field is 2 octets in length, it is used to notify the egress LSR how to choose the return path. The Flag field is a bit vector and has following format:



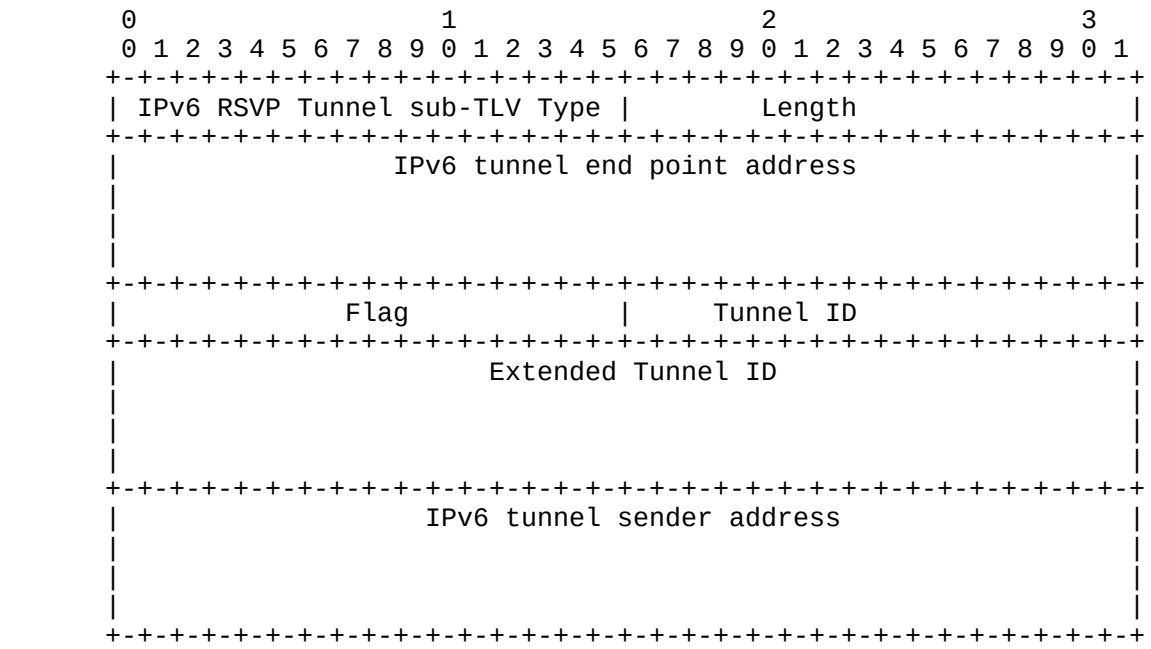
P (Primary): the return path MUST be chosen from the LSPs that have the same Tunnel attributes and the LSP MUST be the primary LSP.

S (Secondary): the return path MUST be chosen from the LSPs that have the same Tunnel attributes and the LSP MUST be the secondary LSP.

P bit and S bit MUST NOT both be set. If P bit and S bit are both not set, the return path could be any one of the LSPs that have the same Tunnel attributes.

3.3.2. IPv6 RSVP Tunnel sub-TLV

The IPv6 RSVP Tunnel sub-TLV is used in the RP TLV to allow the operator to specify a more generic tunnel FEC other than a particular LSP as the return path. The egress LSR chooses an LSP from the LSPs that have the same Tunnel attributes and satisfy the conditions carried in the Flag field. The format of IPv6 RSVP Tunnel sub-TLV is as follows:



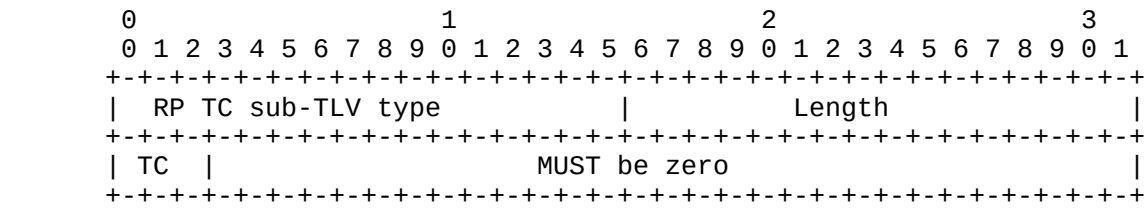
The IPv6 RSVP Tunnel sub-TLV is derived from RSVP IPv6 FEC TLV that is defined in Section 3.2.4 of [RFC4379].All fields have the same semantics as defined in [RFC4379] except that the LSP-ID field is omitted and a new Flag field is defined.

The IPv6 RSVP Tunnel sub-TLV Type field is 2 octets in length, and the recommended type value is 19 (to be confirmed by IANA).

The Flag field is 2 octets in length and is identical to that described in Section 3.3.

3.3.3. RP TC sub-TLV

Reply TOS Byte TLV [RFC4379] is used by the originator of the echo request to request that an echo reply be sent with the IP header TOS byte set to the value specified in the TLV. Similarly, in this document, a new sub-TLV: RP TC sub-TLV is defined and MAY be used by the originator of the echo request to request that an echo reply be sent with the TC bits of the specified return LSP set to the value specified in this sub-TLV. Since there may be more than one FEC sub-TLVs (return paths) specified in the RP TLV, the relevant RP TC sub-TLV MUST directly follow the FEC sub-TLV that identifies the corresponding specified return LSP. The format of RP TC sub-TLV is as follows:



The RP TC sub-TLV Type field is 2 octets in length, and the recommended type value is 17 (to be confirmed by IANA).

The Length field is 2 octets in length, the value of length field is fixed 4.

4. Theory of Operation

The procedures defined in this document currently only apply to "ping" mode. The "traceroute" mode is out of scope for this document.

In [RFC4379], the echo reply is used to report the LSP checking result to the LSP Ping initiator. This document defines a new reply mode and a new TLV (RP TLV) which enable the LSP ping initiator to specify or constrain the return path of the echo reply. Similarly the behavior of echo reply is extended to detect the requested return path by looking at a specified path FEC TLV. This enables LSP Ping to detect failures in both directions of a path with a single operation, this of course cuts in half the operational steps required to verify the end to end bidirectional connectivity and integrity of an LSP.

When the echo reply message is intended to test the return MPLS LSP path(when the A bit and E bit is not set in the previous received echo request message), the destination IP address of the echo reply message MUST never be used in a forwarding decision. To avoid this possibility the destination IP address of the echo reply message that is transmitted along the specified return path MUST be set to numbers from the range 127/8 for IPv4 or 0:0:0:0:0:FFFF:127/104 for IPv6, and the IP TTL MUST be set 1. Of course when the echo reply message is not intended for testing the specified return path (when the A bit or E bit is set in the previous received echo request message) , the procedures defined in [RFC4379] (the destination IP address is copied from the source IP address) apply unchanged.

4.1. Sending an Echo Request

When sending an echo request, in addition to the rules and procedures defined in Section 4.3 of [RFC4379], the reply mode of the echo request MUST be set to "Reply via specified path", and a RP TLV MUST be carried in the echo request message correspondingly. The RP TLV includes one or several reply path sub-TLV(s) to identify the return path(s) the egress LSR should use for its reply.

For a bidirectional LSP, since the ingress LSR and egress LSR of a bidirectional LSP are aware of the relationship between the forward and backward direction LSPs, only the B bit SHOULD be set in the RP TLV. If the operator wants the echo reply to be sent along a different path other than the reverse direction of the bidirectional LSP, "A" bit SHOULD be set or another FEC sub-TLV SHOULD be carried in the RP TLV instead, and the B bit MUST be clear.

In some cases, operators may want to treat two unidirectional LSPs (one for each direction) as a pair. There may not be any binding relationship between the two LSPs. Using the mechanism defined in this document, operators can run LSP Ping one time from one end to complete the failure detection on both unidirectional LSPs. To accomplish this, the echo request message MUST carry (in the RP TLV) a FEC sub-TLV that belongs to the backward LSP.

4.2. Receiving an Echo Request

"Ping" mode processing as defined in Section 4.4 of [RFC4379] applies in this document. In addition, when an echo request is received, if the egress LSR does not know the reply mode defined in this document, an echo reply with the return code set to "Malformed echo request" and the Subcode set to zero will be send back to the ingress LSR according to the rules of [RFC4379]. If the egress LSR knows the reply mode, according to the RP TLV, it SHOULD find and select the desired return path. If there is a matched path, an echo reply with RP TLV that identify the return path SHOULD be sent back to the ingress LSR, the RP return code SHOULD be set to "The echo reply was sent successfully using the specified RP". If there is no such path, an echo reply with RP TLV SHOULD be sent back to the ingress LSR, the RP return code SHOULD be set to relevant code (defined Section 3.2) for the real situation to reflect the result of RP TLV processing and return path selection. For example, if the specified LSP is not found, the egress then chooses another LSP as the return path to send the echo reply, the RP return code SHOULD be set to "The specified RP was not found, the echo reply was sent via other LSP", and if the egress chooses an IP path to send the echo reply, the RP return code SHOULD be set to "The specified RP was not found, the echo reply was sent via IP path". If there is unknown sub-TLV in the received RP

TLV, the RP return code SHOULD be set to "One or more of the sub-TLVs in RP TLV was not understood".

If the A bit of the RP TLV in a received echo request message is set, the egress LSR SHOULD send the echo reply along an non-default return path.

IF the B bit of the RP TLV in a received echo request message is set, the egress LSR SHOULD send the echo reply along the reverse direction of the bidirectional LSP.

If the E bit of the RP TLV in a received echo request message is set, the egress LSR MUST exclude the paths identified by those FEC sub-TLVs carried in the RP TLV and select other path to send the echo reply.

If the A and E bit of the RP TLV in a received echo request message is not set, the echo reply is REQUIRED not only to send along the specified path, but to test the selected return path as well (by carrying the FEC stack information of the return path).

In addition, the FEC validate results of the forward path LSP SHOULD NOT affect the egress LSR continue to test return path LSP.

4.3. Sending an Echo Reply

As described in [RFC4379], the echo reply message is a UDP packet, and it MUST be sent only in response to an MPLS echo request. The source IP address is a routable IP address of the replier, the source UDP port is the well-know UDP port for LSP ping.

When the echo reply is intended to test the return path (both A and E bit are not set in the previous received echo request), the destination IP address of the echo reply message MUST never be used in a forwarding decision. To avoid this problem, the IP destination address of the echo reply message that is transmitted along the specified return path MUST be set to numbers from the range 127/8 for IPv4 or 0:0:0:0:0:FFFF:127/104 for IPv6, and the IP TTL MUST be set to 1. Otherwise, the same as defined in [RFC4379], the destination IP address and UDP port are copied from the source IP address and source UDP port of the echo request.

When sending the echo reply, a RP TLV that identifies the return path MUST be carried, the RP return code SHOULD be set to relevant code that reflects results about how the egress processes the RP TLV in a previous received echo request message and return path selection. By carrying the RP TLV in an echo reply, it gives the Ingress LSR enough information about the reverse direction of the tested path to verify

the consistency of the data plane against control plane. Thus a single LSP Ping could achieve both directions of a path test. If the return path is pure IP path, no sub-TLVs are carried in the RP TLV.

4.4. Receiving an Echo Reply

The rules and process defined in Section 4.6 of [RFC4379] apply here. When an echo reply is received, if the reply mode is "Reply via specified path" and the RP return code is "The echo reply was sent successfully using the specified RP", and if both the A bit and E bit are not set. The ingress LSR MUST do FEC validation (based on the FEC stack information of the return path carried in the RP TLV) as an egress LSR does when receiving an echo request, the FEC validation process (relevant to "ping" mode) defined in Section 4.4.1 of [RFC4379] applies here.

When an echo reply is received with return code set to "Malformed echo request received" and the Subcode set to zero. It is possible that the egress LSR may not know the "Reply via specified path" reply mode, the operator may choose to re-perform another LSP Ping by using one of the four reply modes defined [RFC4379].

On receipt of an echo reply with RP return code in the RP TLV set to "The specified RP was not found, ...", it means that the egress LSR could not find a matched return path as specified. Operators may choose to specify another LSP as the return path or use other methods to detect the path further.

When the LSP Ping initiator fails after some time to receive the echo reply message, the operator MAY initiate another LSP Ping by resending a new echo request carrying a RP TLV with E bit set, the sub-TLVs and/or B bit (when the tested LSP is a bidirectional LSP) identify the previous tried reply paths that are used to notify the egress LSR to send echo reply message along any other workable path other than these failed return paths.

5. Security Considerations

Security considerations discussed in [RFC4379] apply to this document. In addition to that, in order to prevent using the extension defined in this document for "proxying" any possible attacks, the return path LSP MUST have destination to the same node where the forward path is from.

6. IANA Considerations

IANA is requested to assign one new TLV from the "Multiprotocol Label Switching Architecture (MPLS) Label Switched Paths (LSPs) Parameters - TLVs" registry, "TLVs and sub-TLVs" sub- registry; and a set of sub-TLVs under this new TLV; one new Reply Mode from the "Multi-Protocol Label Switching (MPLS) Label Switched Paths (LSPs) Parameters" registry, the "Reply Mode" subregistry.

6.1. New TLV

The IANA is requested to as assign a new TLV from the "Multiprotocol Label Switching Architecture (MPLS) Label Switched Paths (LSPs) Parameters - TLVs" registry, "TLVs and sub-TLVs" sub- registry.

Value	Meaning	Reference
21	Reply Path TLV	this document (sect 3.2)

6.2. Sub-TLVs

Since all existing sub-TLVs and any new sub-TLVs added to the Target FEC Stack TLV apply to the Reply Path TLV, except for the range of 31744-32767 that is left for "Vendor Private Use" in the sub-type space of Target FEC Stack TLV, the sub-TLV space and assignment for Reply Path TLV and Target FEC Stack TLV MUST be kept the same. All new sub-types dedicated added to the Reply Path TLV MUST be assigned from the range of 31744-32767.

6.2.1. New Sub-TLVs

IANA is also requested to assign three new sub-TLV types from "Multiprotocol Label Switching Architecture (MPLS) Label Switched Paths (LSPs) Parameters - TLVs" registry, "TLVs and sub-TLVs" sub-registry for the Reply Path TLV (Type 21).

Sub-type	Value Field	Reference
TBD	RP TC	this document (sect 3.3.3)
TBD	IPv4 RSVP Tunnel	this document (sect 3.3.2)
TBD	IPv6 RSVP Tunnel	this document (sect 3.3.1)

6.2.2. New Reply Mode

IANA is requested to assign a new reply mode code point from the from the "Multi-Protocol Label Switching (MPLS) Label Switched Paths (LSPs) Parameters" registry, the "Reply Mode" subregistry.

Value	Meaning	Reference
5	Reply via specified path	this document (sect 3.1)

7. Contributors

The following individuals also contributed to this document:

Ehud Doron
Orckit-Corrigent

EMail: ehudd@orckit.com

Ronen Solomon
Orckit-Corrigent

EMail: RonenS@orckit.com

Ville Hallivuori
Tellabs
Sinimaentie 6 C
FI-02630 Espoo, Finland

EMail: ville.hallivuori@tellabs.com

Xinchun Guo
EMail: guoxinchun@huawei.com

8. Acknowledgements

The authors would like to thank Adrian Farrel and Peter Ashwood-Smith for their review, suggestion and comments to this document.

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.

[RFC4379] Kompella, K. and G. Swallow, "Detecting Multi-Protocol Label Switched (MPLS) Data Plane Failures", RFC 4379, February 2006.

9.2. Informative References

[RFC3471] Berger, L., "Generalized Multi-Protocol Label Switching (GMPLS) Signaling Functional Description", RFC 3471, January 2003.

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

[RFC3945] Mannie, E., "Generalized Multi-Protocol Label Switching (GMPLS) Architecture", RFC 3945, October 2004.

[RFC5654] Niven-Jenkins, B., Brungard, D., Betts, M., Sprecher, N., and S. Ueno, "Requirements of an MPLS Transport Profile", RFC 5654, September 2009.

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

[RFC5884] Aggarwal, R., Kompella, K., Nadeau, T., and G. Swallow, "Bidirectional Forwarding Detection (BFD) for MPLS Label Switched Paths (LSPs)", RFC 5884, June 2010.

Authors' Addresses

Mach(Guoyi) Chen
Huawei Technologies Co., Ltd
Q14 Huawei Campus, No. 156 Beiqing Road, Hai-dian District
Beijing 100095
China

Email: mach@huawei.com

Wei Cao
Huawei Technologies Co., Ltd
Q14 Huawei Campus, No. 156 Beiqing Road, Hai-dian District
Beijing 100095
China

Email: wayne.caowei@huawei.com

So Ning
Verizon Inc.
2400 N. Glenville Rd.,
Richardson, TX 75082
USA

Email: ning.so@verizonbusiness.com

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

Email: frederic.jounay@orange-ftgroup.com

Simon Delord
Alcatel-Lucent
Building 3, 388 Ningqiao Road, Jinqiao, Pudong
Shanghai 201206
China

Email: simon.delord@alcatel-lucent.com

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: May 3, 2012

R. Winter, Ed.
NEC
E. Gray, Ed.
Ericsson
H. van Helvoort
Huawei Technologies Co., Ltd.
M. Betts
ZTE

October 31, 2011

MPLS-TP Identifiers Following ITU-T Conventions
draft-ietf-mpls-tp-itu-t-identifiers-02

Abstract

This document specifies an extension to the identifiers to be used in the Transport Profile of Multiprotocol Label Switching (MPLS-TP). Identifiers that follow IP/MPLS conventions have already been defined. This memo augments that set of identifiers for MPLS-TP management and OAM functions to include identifier information in a format typically used 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 May 3, 2012.

Table of Contents

1. Introduction

This document augments the initial set of identifiers to be used in the Transport Profile of Multiprotocol Label Switching (MPLS-TP) specified in RFC 6370 [RFC6370].

RFC 6370 [RFC6370] defines a set of MPLS-TP transport and management entity identifiers to support bidirectional (co-routed and associated) point-to-point MPLS-TP LSPs, including PWs and Sections which follow the IP/MPLS conventions.

This document specifies an alternative way to uniquely identify an operator/service provider based on ITU-T conventions and specifies how this operator/service provider identifier can be used to make the existing set of MPLS-TP transport and management entity identifiers, defined by RFC 6370 [RFC6370], globally unique.

This document solely defines those identifiers. Their use and possible protocols extensions to carry them is out of scope in this document.

In this document, we follow the notational convention laid out in RFC 6370 [RFC6370].

1.1. Terminology

CC: Country Code

ICC: ITU-T Carrier Code

ITU-T: International Telecommunication Union Telecommunication Standardization Sector

LSP: Label Switched Path

MEG: Maintenance Entity Group

MEP: Maintenance Entity Group End Point

MPLS: Multi-Protocol Label Switching

PW: Pseudowire

TSB: (ITU-T) Telecommunication Standardization Bureau

UMC: Unique MEG ID Code

1.2. Requirements notation

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

2. Uniquely Identifying an Operator - the ICC_Operator_ID

In RFC 6370 [RFC6370] an operator is uniquely identified by the Global_ID which is based on the AS number of the operator. The ITU-T however traditionally identifies operators/service providers based on the ITU-T Carrier Code (ICC) as specified in [M1400].

The ITU-T Telecommunication Standardization Bureau (TSB) maintains a list of assigned ICCs [ICC-list]. Note that ICCs can be assigned to both, ITU-T members as well as non-members, all of which are referenced at [ICC-list]. The national regulatory authorities act as an intermediary between the ITU/TSB and operators/service providers. Amongst the things that the national authorities are responsible for in the process of assigning an ICC is to ensure that the Carrier Codes are unique within their country.

The ICC itself is a string of one to six characters, each character being either alphabetic (i.e. A-Z) or numeric (i.e. 0-9). Alphabetic characters in the ICC SHOULD be represented with upper case letters.

Global uniqueness is assured by concatenating the ICC with a Country Code (CC). The Country Code (alpha-2) is a string of two alphabetic characters represented with upper case letters (i.e., A-Z). The Country Code format is defined in ISO 3166-1 [ISO3166-1]. Together, the CC and the ICC form the ICC_Operator_ID as CC::ICC.

3. Use of the ICC_Operator_ID

The ICC_Operator_ID is used as a replacement for the Global_ID as specified in RFC 6370 [RFC6370], i.e. its purpose is to provide a globally unique context for other MPLS-TP identifiers.

As an example, an Interface Identifier (IF_ID) in RFC 6370 [RFC6370] is specified as the concatenation of the Node_ID (a unique 32-bit value assigned by the operator) and the Interface Number (IF_Num, a 32-bit unsigned integer assigned by the operator that is unique within the scope of a Node_ID). To make this IF_ID globally unique the Global_ID is prefixed. This memo specifies the ICC_Operator_ID as an alternative format which, just like the Global_ID, is prefixed

to the IF_ID. Using the notation from RFC 6370 [RFC6370]:

Global_ID::Node_ID::IF_Num

is functionally equivalent to:

ICC_Operator_ID::Node_ID::IF_Num

The same substitution procedure applies to all identifiers specified in RFC 6370 [RFC6370] except for the other alternatives mentioned in this document.

4. ICC_Operator_ID-based MEG Identifiers

The ITU-T format of MEG_IDs for MPLS-TP Sections, LSPs and Pseudowires is based on the globally unique ICC_Operator_ID. In this case, the MEG_ID is a string of up to 15 characters. It consists of three subfields: the Country Code (as described in Section 2), the ICC (as described in Section 2) which together form the ICC_Operator_ID, followed by a Unique MEG ID Code (UMC) as defined in [Y.1731_cor1].

The resulting MEG_ID therefore looks like the following:

CC:ICC:UMC

To avoid the potential for a short (i.e. less than 6 Character) ICC code in combination with a UMC not being unique the UMC MUST start with a special character that is not allowed in the ICC such as the "/" character. A side effect of this is that the MEG_ID can be decomposed into its individual components by a receiver.

The UMC MUST be unique within the organization identified by the combination of CC and ICC.

The ICC_Operator_ID-based MEG_ID may be applied equally to a single MPLS-TP Section, LSP or Pseudowire.

5. ICC_Operator_ID-based MEP Identifiers

ICC_Operator_ID-based MEP_IDs for MPLS-TP LSPs and Pseudowires are formed by appending a 16-bit index to the MEG_ID defined in Section 4 above. Within the context of a particular MEG, we call the identifier associated with a MEP the MEP Index (MEP_Index). The MEP_Index is administratively assigned. It is encoded as a 16-bit unsigned integer and MUST be unique within the MEG. An

ICC_Operator_ID-based MEP_ID is structured as:

MEG_ID::MEP_Index

An ICC_Operator_ID-based MEP ID is globally unique by construction given the ICC_Operator_ID-based MEG_ID's global uniqueness.

6. Security Considerations

This document extends an existing information model and, as such, does in itself not introduce new security concerns. But, as mentioned in the security considerations section of the document that is being augmented, protocol specifications that describe use of this information model may introduce security risks and concerns about authentication of participants. For this reason, these protocol specifications need to describe security and authentication concerns that may be raised by the particular mechanisms defined and how those concerns may be addressed.

7. IANA Considerations

There are no IANA actions resulting from this document.

8. References

8.1. Normative References

[ISO3166-1] "Codes for the representation of names of countries and their subdivisions -- Part 1: Country codes", ISO 3166-1.

[M1400] "Designations for interconnections among operators' networks", ITU-T Recommendation M.1400, July 2006, <http://www.itu.int/rec/T-REC-M.1400-200607-I/en>.

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

[RFC6370] Bocci, M., Swallow, G., and E. Gray, "MPLS Transport Profile (MPLS-TP) Identifiers", RFC 6370, September 2011.

[Y.1731_cor1] "OAM functions and mechanisms for Ethernet based networks - Corrigendum 1", ITU-T Recommendation ITU-T G.8013/Y.1731 (2011) Corrigendum 1.

8.2. Informative References

[ICC-list]
"List of ITU Carrier Codes (ICCs)",
<http://www.itu.int/oth/T0201>.

Authors' Addresses

Rolf Winter (editor)
NEC

Email: rolf.winter@neclab.eu

Eric Gray (editor)
Ericsson

Email: eric.gray@ericsson.com

Huub van Helvoort
Huawei Technologies Co., Ltd.

Email: huub.van.helvoort@huawei.com

Malcolm Betts
ZTE

Email: malcolm.betts@zte.com.cn

Internet Engineering Task Force
Internet-Draft
Intended status: Informational
Expires: April 30, 2012

L. Fang, Ed.
Cisco Systems
B. Niven-Jenkins, Ed.
Velocix
S. Mansfield, Ed.
Ericsson
R. Graveman, Ed.
RFG Security
October 31, 2011

MPLS-TP Security Framework
draft-ietf-mpls-tp-security-framework-02

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, a 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 and potential security threats and mitigation 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 April 30, 2012.

Copyright Notice

Table of Contents

1. Introduction

1.1. Background and Motivation

This document provides a security framework for Multiprotocol Label Switching Transport Profile (MPLS-TP).

The 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 and the fast-growing bandwidth demand accelerated by new packet-based services and multimedia applications, from Ethernet Services, Layer 2 and Layer 3 VPNS, and triple play to Mobile Access Network (RAN) backhaul, etc. MPLS-TP is based on MPLS technologies to take advantage of this technology's maturity, and MPLS-TP is required to maintain the transport characteristics of MPLS.

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 Systems (NMS) or Operation Support Systems (OSS). There are also needs for MPLS-TP and MPLS interworking.

The security aspects for the new extensions particularly designed for MPLS-TP need to be addressed. The security models, requirements, threats, and defense techniques previously defined in [RFC5921] can be applied to reuse existing functionalities in MPLS and GMPLS but are 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 specific to MPLS-TP. It provides the security requirements for MPLS-TP, defines security models that apply to various MPLS-TP deployment scenarios, and identifies 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, because these 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], and the content of [RFC5920] that has no new impact on MPLS-TP is not 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 operations security considerations. This document does not define the specific mechanisms or methods that must be implemented to satisfy the security requirements.

The issues and areas addressed with respect to MPLS-TP security are:

- G-Ach (control plane attack, DoS attack, message intercept, etc.)
- ID Spoofing
- Loopback attacks
- NMS attacks
- NMS and CP interaction vulnerabilities
- MIP and MEP assignment and attacks on these mechanisms
- Topology discovery vulnerabilities
- Data plane authentication
- Label authentication
- DoS attacks on the Data Plane
- Performance Monitoring vulnerabilities

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].

- BFD: Bidirectional Forwarding Detection
- CE: Customer-Edge device
- DoS: Denial of Service
- DDoS: Distributed Denial of Service
- GAL: Generic Alert Label
- G-ACH: Generic Associated Channel
- GMPLS: Generalized Multi-Protocol Label Switching
- LDP: Label Distribution Protocol
- LSP: Label Switched Path
- MCC: Management Communication Channel
- MEP: Maintenance End Point
- MIP: Maintenance Intermediate Point
- MPLS: MultiProtocol Label Switching
- OAM: Operations, Administration, and Management
- PE: Provider-Edge device
- PSN: Packet-Switched Network
- PW: Pseudowire
- RSVP: Resource Reservation Protocol
- RSVP-TE: Resource Reservation Protocol with Traffic Engineering Extensions
- S-PE: Switching Provider Edge

- SSH: Secure Shell
- TE: Traffic Engineering
- TLS: Transport Layer Security
- T-PE: Terminating Provider Edge
- VPN: Virtual Private Network
- WG: Working Group of IETF
- 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 and mitigation techniques and procedures

2. Security Reference Models

This section defines reference models 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.

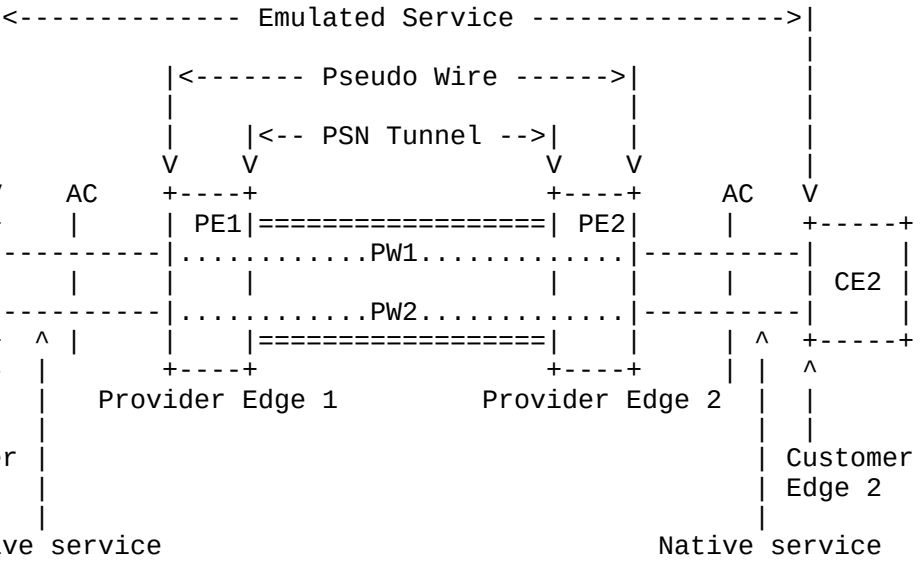
This document defines a trusted zone as being where a single SP has total operational control over that part of the network. A primary concern is about security aspects that relate to breaches of security from the "outside" of a trusted zone to the "inside" of this zone.

2.1. Security Reference Model 1

In reference model 1, a single SP has total control of the PE/T-PE to PE/T-PE part 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).



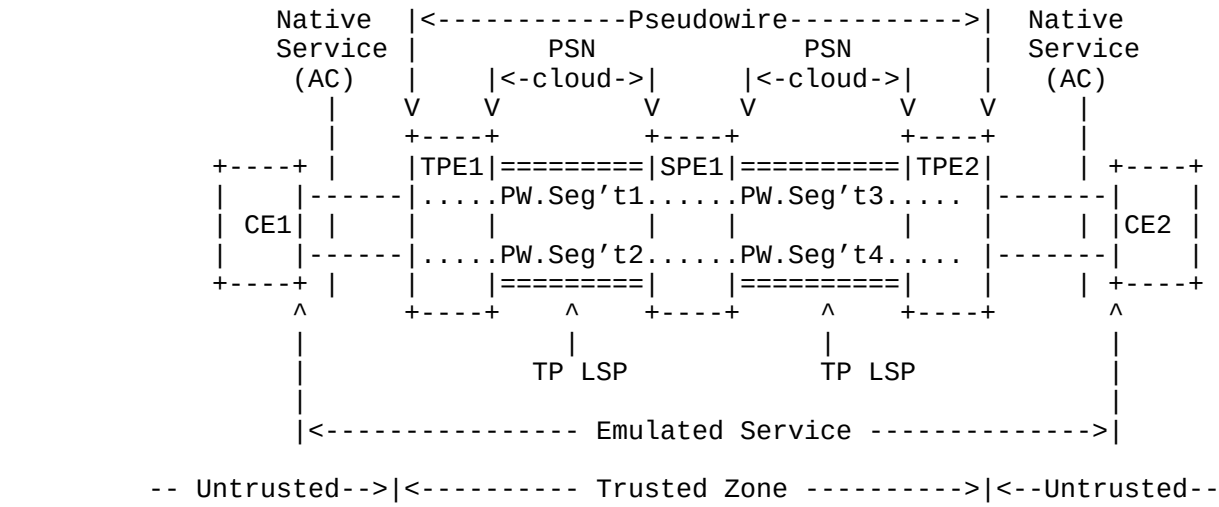
MPLS-TP Security Model 1 (a)

Figure 1

AC: Attachment Circuit

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)

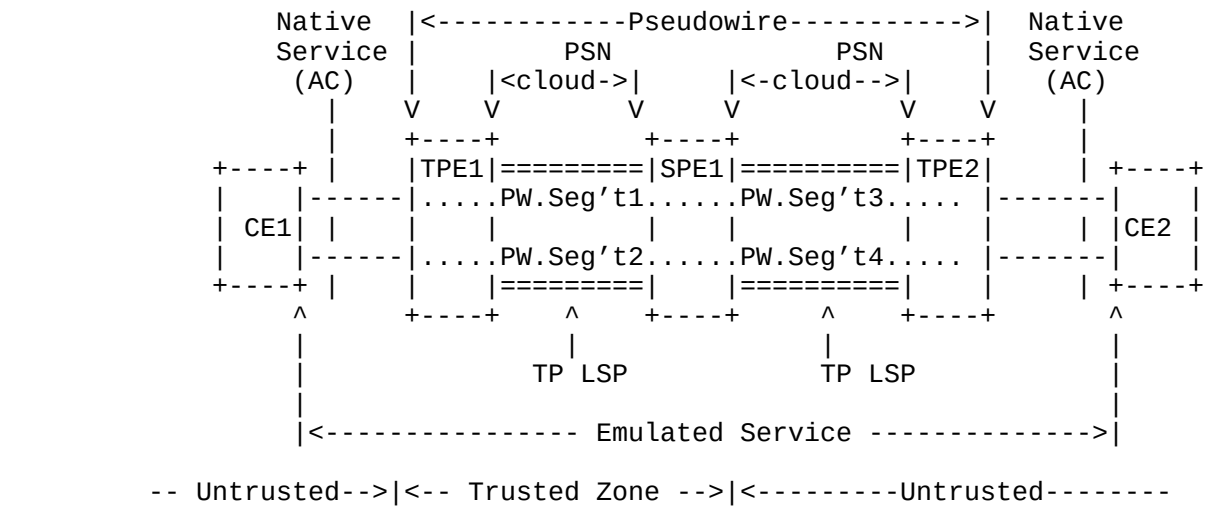
Figure 2

2.2. Security Reference Model 2

In reference model 2, a single SP does not have total control of the PE/T-PE to PE/T-PE part 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).

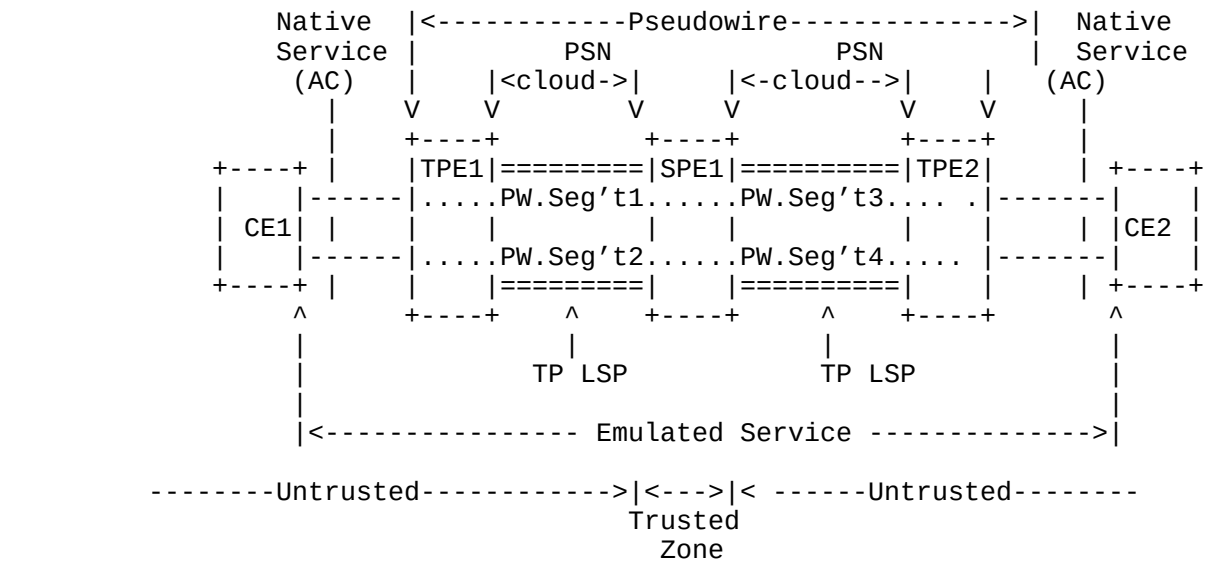


MPLS-TP Security Model 2 (a)

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).

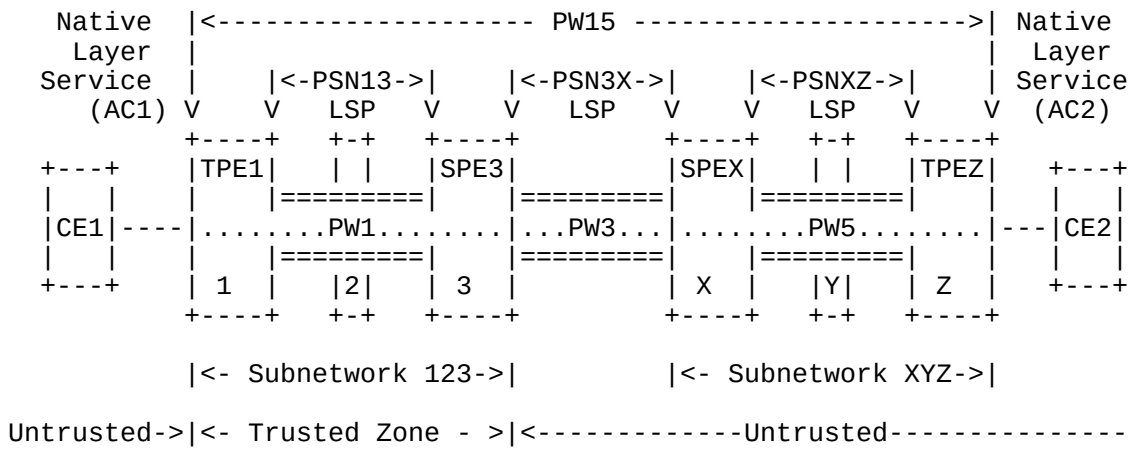


MPLS-TP Security Model 2 (b)

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).

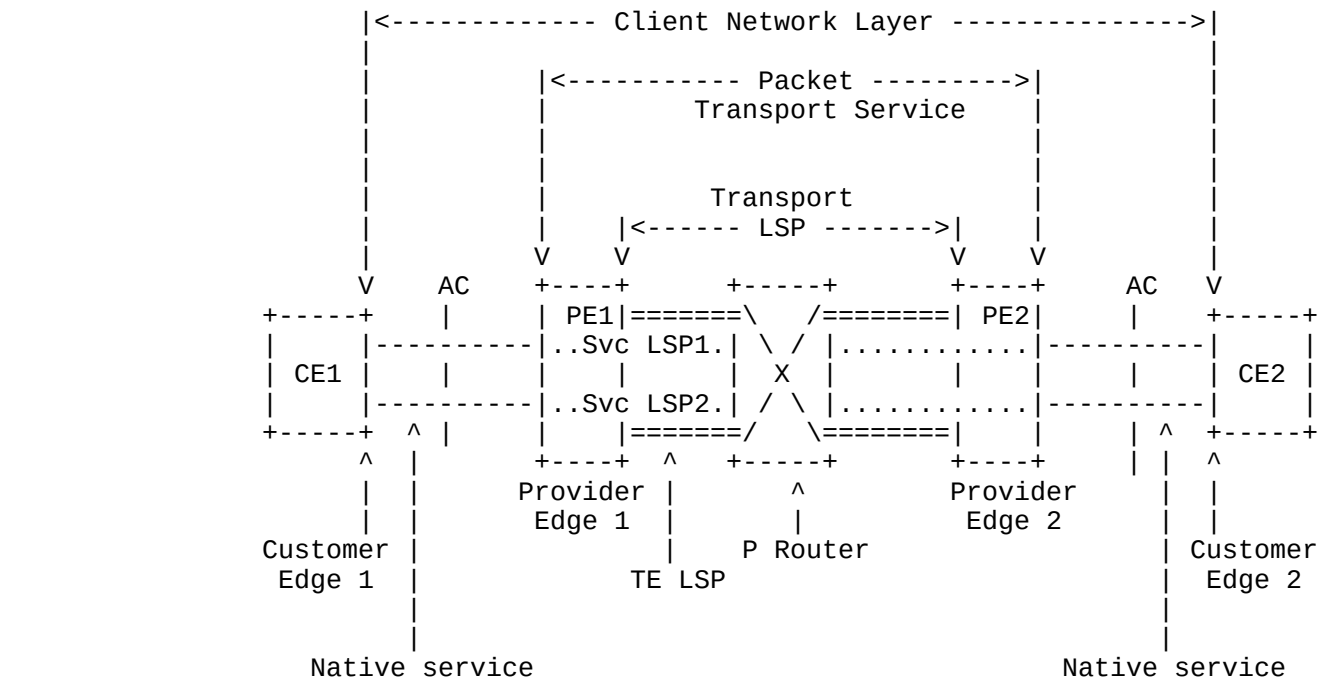


MPLS-TP Security Model 2 (c)

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)

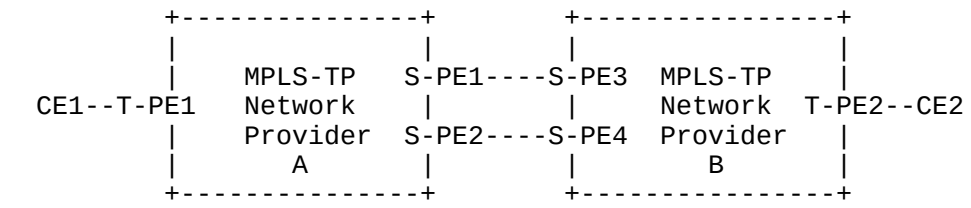
Figure 6

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 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 a trusted zone MUST NOT be compromised by interconnecting one SP's MPLS-TP or MPLS infrastructure with another SP's core devices, 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 an MPLS-TP network infrastructure. The MPLS-TP network can be operated without a control plane or via dynamic control plane protocols. The security requirements related to new MPLS-TP OAM, recovery mechanisms, MPLS-TP and MPLS interconnection, and MPLS-TP specific operations are addressed in this section.

A service provider may choose the deployment options best fitting for its network operations. This document does not mandate that an MPLS-TP 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 the data plane from the control plane and the management plane. That is, if the control plane, management plane, or both 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 LSPs and PWs with or without an NMS or 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 not in the trusted zone.

- MPLS-TP MUST support non-IP path options in addition to the IP loopback option. Non-IP path options when used in security model 2 (Section 2.2) may help to lower the potential risk of attacks on the S-PE/T-PE in the trusted zone.

- 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.

- MPLS-TP MUST support mechanisms to prevent Denial of Service (DoS) attacks via any in-band OAM or G-ACh/GAL.

- MPLS-TP MUST support hiding of the Service Provider’s infrastructure for all reference models regardless of whether the network(s) are using static configuration or a dynamic control plane.

- Management security requirements from [RFC5951] include the following:

* MPLS-TP MUST support security for the management communication channel (MCC).

* 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 provide support for confidentiality and data integrity protection for applications.

* The MCC MUST support the use of a flexible set of strong, open, and standard cryptographic algorithms (see Section 2.2 of [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 potential and impact for DoS and DDoS attacks by using secure protocols, turning off unnecessary processes, logging and monitoring, and ingress filtering. See [RFC4732] for background on DoS in the context of the Internet.

o MPLS-TP MUST provide protection from operational errors. Due to the extensive use of static provisioning with or without a NMS or OSS, the prevention of configuration errors should be addressed as a major set of 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 inauthentic 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 or BFD messages:

    a. GAL label or BFD label manipulation, which includes insertion of false labels or messages, and modification or removal of GAL labels or messages by attackers.

    b. DoS attack through in-band OAM G-ACH/GAL and BFD messages.

3. Disruption of a provider's or user's connectivity, or degradation of a provider's service quality.

    a. Attacks against provider connectivity:

        + In the case in which an NMS is used for LSP set-up, the attacks occur through attacks on the NMS.

        + In the case in which dynamic provisioning is used, the attacks occur on the dynamic control plane. Most aspects of these are addressed in [RFC5920].

    b. Attacks against user connectivity. These are similar to PE/CE attacks against access in typical MPLS networks and are addressed in [RFC5920].

4. Probing a provider's network to determine its configuration, capacity, or usage. These types of attack can occur through attacks against an NMS in the case of static provisioning, or through attacks against the control plane in dynamic MPLS networks. They 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:

- Other users whose services are provided by the same MPLS-TP core.
- The MPLS-TP SP or persons working for it.
- Other persons who obtain physical access to a MPLS-TP SP's site.
- Other persons who use social engineering methods to influence the behavior of a SP's personnel.
- Users of the MPLS-TP network itself.
- Others, e.g., attackers from the other sources, including the Internet if connected.
- Other SPs in the case of MPLS-TP inter-provider connection. The provider may or may not be using MPLS-TP.
- Those who create, deliver, install, and maintain hardware or 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:

- A MPLS-TP network inter-connecting with another provider's core
- 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. Even though surveys show that 40% to 60% of attacks originate from insiders, 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

- MPLS-TP LSP creation by an unauthorized element
- LSP message interception
- Attacks on G-Ach
- Attacks against LDP
- Attacks against RSVP-TE
- Attacks against GMPLS
- Denial of Service Attacks on the Network Infrastructure
- Attacks on the SP's MPLS/GMPLS Equipment via Management Interfaces
- Social Engineering Attacks on the SP's Infrastructure
- Cross-Connection of Traffic between Users
- Attacks against Routing Protocols
- 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.

- Unauthorized Observation of Data Traffic
- Modification of Data Traffic
- Insertion of Inauthentic Data Traffic: Spoofing and Replay
- Unauthorized Deletion of Data Traffic
- 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 specific operational environment determines the security requirements for any instance of MPLS-TP. Therefore, protocol designers should provide a full set of security capabilities, which can be selected and used where appropriate. 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 techniques discussed here include entity authentication for identity verification, encryption for confidentiality, message integrity and replay detection to ensure the validity of message streams, network-based access controls such as packet filtering and firewalls, host-based access controls, isolation, aggregation, and event logging. Where these techniques apply to MPLS and GMPLS in general, they are described in Section 5.2 of [RFC5920]. The remainder of this section covers aspects that apply particularly to MPLS-TP.

5.1. Authentication

To prevent security issues arising from impersonation, masquerade, or some DoS attacks or from malicious or accidental misconfiguration, it is critical that MPLS-TP devices should accept connections or control messages only from known sources. Authentication refers to methods for ensuring that the identities of message sources are properly verified by the MPLS-TP devices with which they communicate. This section focuses on 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.

This type of authentication should be bi-directional. The PE or CE needs 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 (e.g., 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 variety of public-private key systems. Some of these use digital certificates binding a user's name and public key. One method of using digital certificates is within a hierarchical Certification Authority system.

5.2. Access Control Techniques

Many of the security issues related to management interfaces can be addressed through the use of authentication as described in Section 5.1. However, additional security may be provided by controlling access to management interfaces or to specific resources with an access control model. In addition to identification and authentication, access control deals with authorization.

Much of the work on security for SNMP has focused on access control models. For the most recent version of SNMP security, see the work of the ISMS WG.

The Optical Internetworking Forum has done relevant work on Protecting interfaces to management systems with TLS, SSH, 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].

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

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 networks and services may be subject to attacks from a variety of security threats. Many types of threats are described in the Security Requirements (Section 3) Section of this document. The defensive techniques described in this document and elsewhere provide significant levels of protection from many of these 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.

Also, extensive logging of normal processing, error conditions and security events can be an invaluable source of information for tracking down attacks, recovering from them, and determining how to prevent future attacks. Different methods may be appropriate from case to case, and in fact comparing the same or similar information obtained in different ways (e.g., with syslog and SNMP) has sometimes reveals subtle security flaws or actual intrusions. Implementations should also pay attention to the security of the logs themselves.

7. Security Considerations

Security considerations constitute the sole subject of this document and hence are discussed throughout.

The document describes a variety of defensive techniques that may be used to counter the potential 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.

Note: A single document updating these two OIF agreements may be published in November 2011. If so, it will be posted at http://www.oiforum.com/public/impagreements.html.

[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
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

Richard F. Graveman (editor)
RFG Security, LLC
15 Park Avenue
Morristown, NJ 07960 US

Email: rfg@acm.org

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/1id-abstracts.txt.

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

This Internet-Draft will expire on December 17, 2011.

Table of Contents

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.

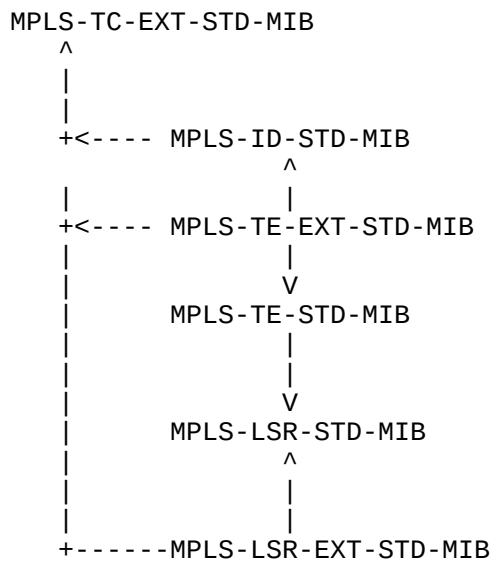
6.4.mplsTunnelExtTable

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.

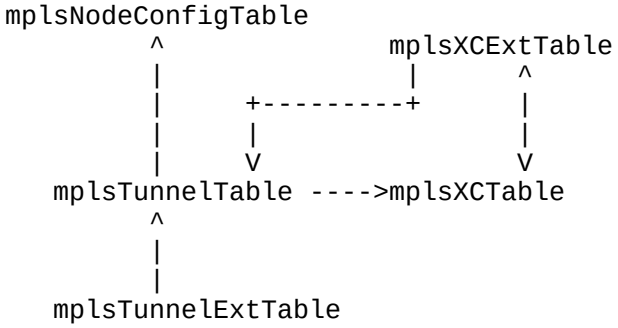
MPLS-LSR-STD-MIB [RFC 3813] is extended by MPLS-LSR-EXT-STD-MIB mib module for pointing back to the tunnel entry for easy tunnel access from XC entry.

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,
         mplsXCLspId               = 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 AIIs.
            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 principles (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

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: February 7, 2012

L. Jin
ZTE
F. Jounay
France Telecom
I. Wijnands
Cisco Systems
N. Leymann
Deutsche Telekom AG
August 6, 2011

Multicast LDP extension for hub & spoke multipoint LSP
draft-jin-jounay-mpls-mldp-hsmp-04.txt

Abstract

This draft introduces a hub & spoke multipoint LSP (short for HSMP LSP), which allows traffic both from root to leaf through P2MP LSP and also leaf to root along the co-routed reverse path. That means traffic entering the HSMP LSP from application/customer at the root node travels downstream, exactly as if it was traveling downstream along a P2MP LSP to each leaf node, and traffic entering the HSMP LSP at any leaf node travels upstream along the tree to the root. A packet traveling upstream should be thought of as being unicast to the root, except that it follows the path of the tree rather than ordinary unicast path.

Requirements Language

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

Status of this Memo

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

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

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

This Internet-Draft will expire on February 7, 2012.

Table of Contents

1. Introduction

The point-to-multipoint LSP defined in [I-D.ietf-mpls-ldp-p2mp] allows traffic to transmit from root to several leaf nodes, and multipoint-to-multipoint LSP allows traffic from every node to transmit to every other node. This draft introduces a hub & spoke multipoint LSP (short for HSMP LSP), which allows traffic both from root to leaf through P2MP LSP and also leaf to root along the co-routed reverse path. That means traffic entering the HSMP LSP at the root node travels downstream, exactly as if it was traveling downstream along a P2MP LSP, and traffic entering the HSMP LSP at any other node travels upstream along the tree to the root. A packet traveling upstream should be thought of as being unicast to the root, except that it follows the path of the tree rather than ordinary unicast path.

2. Applications

In some cases, the P2MP LSP may not have a reply path for the OAM message (e.g, LSP Ping). If P2MP LSP is provided by HSMP LSP, then the upstream path could be exactly used as the OAM message reply path. This is especially useful in the case of P2MP LSP fault detection, performance measurement, root node redundancy and etc. There are several other applications that could take advantage of such kind of LDP based HSMP LSP as described below.

2.1. Time synchronization

The delivery of time synchronization to end equipments, such as base stations, can be achieved using a time protocol as [IEEE1588v2] (also known as PTP). This protocol defines Transparent Clock (TC) function, which can be used in transport nodes to improve the accuracy of time synchronization. Two types of TCs exist in [IEEE1588v2]: End-to-end Transparent Clock (E2E TC) and Peer-to-peer Transparent Clock (P2P TC). P2P TCs assume that the link delays between the different nodes are calculated.

Assuming that a chain of P2P TCs is used between a PTP master and a PTP slave, time synchronization can be delivered to the PTP slave by sending timestamps only in the direction master to slave (one way mode), via PTP Sync messages. This is possible thanks to the link delay calculation performed locally by each node, which enable to calculate the propagation delay over the path. This scenario permits that the same PTP Sync messages would be sent by the PTP master to all the PTP slaves.

In this scenario (chain of P2P TCs), the PTP slave might have to send

also messages (not carrying timestamps) back to the PTP master in some cases. For instance, PTP Signaling messages could be sent back to the PTP master. These PTP Signaling messages are not intended to be received by the other PTP slaves.

IEEE1588 over MPLS is defined in [I-D.ietf-tictoc-1588overmpls]. By using point-to-multipoint technology to transmit PTP Sync messages will greatly improve the bandwidth usage for above applications. Unfortunately current point-to-multipoint LSP only provides unidirectional path from source to leaf, which cannot fulfill the above new requirement (i.e. need for a reverse path for the PTP Signaling messages). The main motivation of this draft is to solve the new problem. LDP based HSMP LSP described in this draft provides co-routed reverse path from leaf to root based on current unidirectional point-to-multipoint LSP.

There are two main specific scenarios for timing synchronization based on [IEEE1588v2]:1. HSMP for phase/time delivery with TCKs.2. HSMP for phase/time delivery with BCKs.The benefit of using mLDP based HSMP LSP here is to provision dynamically the topology.

Time synchronization is also required for accurate quantification of one-way delay as described in [I-D.ietf-mpls-loss-delay]. HSMP LSP can be used to do time synchronization based on [IEEE1588v2] with a chain of P2P TCs for P2MP LSP or P2MP PW.

2.2. IPTV scenario

The mLDP based HSMP LSP can also be applied in a typical IPTV scenario. There is usually only one location with senders but there are many receiver locations. If IGMP is used for signaling between senders and receivers, the messages from the receivers are travelling only from the leaves to the root (and from root towards leaves) but not from leaf to leaf. In addition traffic from the root is only replicated towards the leaves. Then leaf node receiving IGMP message (for SSM case) will join HSMP LSP, and send IGMP message upstream to root along HSMP LSP.

2.3. P2MP PW based L2VPN (VPMS and VPLS)

Point to multipoint PW described in [I-D.ietf-pwe3-p2mp-pw] requires to setup reverse path from leaf node (referred as egress PE) to root node (referred as ingress PE), if HSMP LSP is used to multiplex P2MP PW, the reverse path can also be multiplexed to HSMP upstream path to avoid setup independent reverse path. In that case, the operational cost will be reduced for maintaining only one HSMP LSP, instead of P2MP LSP and n (number of leaf nodes) P2P reverse LSPs.

The VPMS defined in [I-D.ietf-l2vpn-vpms-frmwk-requirements] requires reverse path from leaf to root node. The P2MP PW multiplexed to HSMP LSP can provide VPMS with reverse path, without introducing independent reverse path from each leaf to root.

The P2MP PW multiplexed to HSMP LSP can also be used to VPLS [RFC4762], which will reduce the overall broadcast/multicast utilization on a VPLS. In current VPLS implementations with a full mesh of P2P PWs between PEs, broadcast, multicast and unknown traffic are not efficiently propagated on the physical links between PEs and Ps.

In the VPLS implementation scenario with P2MP PW multiplexed to HSMP LSPs, each PE signals a P2MP PW with itself as a root to all other PEs in the VPLS. Thereafter, all broadcast/multicast/unknown traffic from this PE will use this P2MP PW multiplexed to HSMP. Unicast traffic from a particular PE (e.g. PE1) to another PE (e.g. PE2) will be sent from leaf to root using the reverse path of P2MP PW where PE2 is the root.

This simplifies the VPLS implementation by: 1. reducing traffic utilization from broadcast, multicast and unknown traffic; 2. reducing the total number of LSPs maintained by each PE (i.e. instead of requiring a full mesh of PW, now only require one P2MP PW multiplexed to HSMP per PE).

3. Terminology

mLDP: Multicast LDP.

P2MP LSP: An LSP that has one Ingress LSR and one or more Egress LSRs.

MP2MP LSP: An LSP that connects a set of nodes, such that traffic sent by any node in the LSP is delivered to all others.

HSMP LSP: hub & spoke multipoint LSP. An LSP allows traffic both from root to leaf through P2MP LSP and also leaf to root along the co-routed reverse path.

4. Setting up HSMP LSP with LDP

HSMP LSP is similar with MP2MP LSP described in [I-D.ietf-mpls-ldp-p2mp], with the difference that the leaf LSRs can only send traffic to root node along the same path of traffic from root node to leaf node.

HSMP LSP consists of a downstream path and upstream path. The downstream path is same as MP2MP LSP, while the upstream path is only from leaf to root node, without communication between leaf and leaf nodes. The transmission of packets from the root node of a HSMP LSP to the receivers is identical to that of a P2MP LSP. Traffic from a leaf node follows the upstream path toward the root node, along the identical path of downstream path.

For setting up the upstream path of a HSMP LSP, ordered mode MUST be used which is same as MP2MP. Ordered mode can guarantee a leaf to start sending packets to root immediately after the upstream path is installed, without being dropped due to an incomplete LSP.

Due to much of same behavior between HSMP LSP and MP2MP LSP, the following sections only describe the difference between the two entities.

4.1. Support for HSMP LSP setup with LDP

HSMP LSP also needs the LDP capabilities [RFC5561] to indicate the supporting for the setup of HSMP LSPs. An implementation supporting the HSMP LSP procedures specified in this document MUST implement the procedures for Capability Parameters in Initialization Messages. Advertisement of the HSMP LSP Capability indicates support of the procedures for HSMP LSP setup.

A new Capability Parameter TLV is defined, the HSMP LSP Capability. Following is the format of the HSMP LSP Capability Parameter.

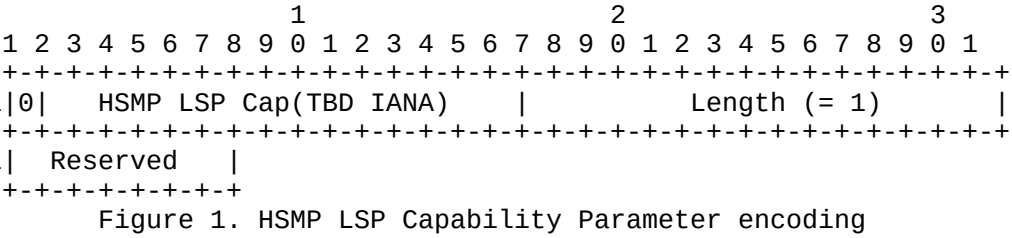


Figure 1. HSMP LSP Capability Parameter encoding

The HSMP LSP capability type is to be assigned by IANA.

4.2. HSMP FEC Elements

Similar as MP2MP LSP, we define two new protocol entities, the HSMP downstream FEC and upstream FEC Element. Both elements will be used as FEC Elements in the FEC TLV. The structure, encoding and error handling for the HSMP downstream and upstream FEC Elements are the

same as for the MP2MP FEC Element described in [I-D.ietf-mpls-ldp-p2mp] Section 4.2. The difference is that two additional new FEC types are used: HSMP downstream type (TBD, IANA) and HSMP upstream type (TBD, IANA).

4.3. Using the HSMP FEC Elements

In order to describe the message processing clearly, following defines the processing of the HSMP FEC Elements, which is inherited from [I-D.ietf-mpls-ldp-p2mp] section 4.3.

1. HSMP downstream LSP <X, Y> (or simply downstream <X, Y>): a HSMP LSP downstream path with root node address X and opaque value Y.

2. HSMP upstream LSP <X, Y> (or simply upstream <X, Y>): a HSMP LSP upstream path for root node address X and opaque value Y which will be used by any of downstream node to send traffic upstream to root node.

3. HSMP downstream FEC Element <X, Y>: a FEC Element with root node address X and opaque value Y used for a downstream HSMP LSP.

4. HSMP upstream FEC Element <X, Y>: a FEC Element with root node address X and opaque value Y used for an upstream HSMP LSP.

5. HSMP-D Label Map <X, Y, L>: A Label Map message with a single HSMP downstream FEC Element <X, Y> and label TLV with label L. Label L MUST be allocated from the per-platform label space of the LSR sending the Label Map Message.

6. HSMP-U Label Map <X, Y, Lu>: A Label Map message with a single HSMP upstream FEC Element <X, Y> and label TLV with label Lu. Label Lu MUST be allocated from the per-platform label space of the LSR sending the Label Map Message.

4.3.1. HSMP LSP Label Map

This section specifies the procedures for originating HSMP Label Map messages and processing received HSMP label map messages for a particular HSMP LSP. The procedure of downstream HSMP LSP is same as that of downstream MP2MP LSP described in [I-D.ietf-mpls-ldp-p2mp]. Under the operation of ordered mode, the upstream LSP will be setup by sending HSMP LSP mapping message with label which is allocated by upstream LSR to its downstream LSR one by one from root to leaf node, installing the upstream forwarding table by every node along the LSP. Detail procedure of upstream HSMP LSP is different with that of upstream MP2MP LSP, and is specified in below section.

All labels discussed here are downstream-assigned [RFC5332] except those which are assigned using the procedures described in section 5.

Determining the upstream LSR for a HSMP LSP <X, Y> follows the procedure for a MP2MP LSP described in [I-D.ietf-mpls-ldp-p2mp] Section 4.3.1.1.

Determining one's downstream HSMP LSR procedure is much same as defined in [I-D.ietf-mpls-ldp-p2mp] section 4.3.1.2. A LDP peer U which receives a HSMP-D Label Map from a LDP peer D will treat D as downstream HSMP LSR.

Determining the forwarding interface to an LSR has same procedure as defined in [I-D.ietf-mpls-ldp-p2mp] section 2.4.1.2.

4.3.1.1. HSMP LSP leaf node operation

The leaf node operation is same as the operation of MP2MP LSP defined in [I-D.ietf-mpls-ldp-p2mp] section 4.3.1.4, only with different FEC element processing and specified below.

A leaf node Z will send a HSMP-D Label Map <X, Y, L> to U, instead of MP2MP-D Label Map <X, Y, L>. and expects a HSMP-U Label Map <X, Y, Lu> from node U and checks whether it already has forwarding state for upstream <X, Y>. The created forwarding state on leaf node Z is same as the leaf node of MP2MP LSP. Z will push label Lu onto the traffic that Z wants to forward over the HSMP LSP.

4.3.1.2. HSMP LSP transit node operation

Suppose node Z receives a HSMP-D Label Map <X, Y, L> from LSR D, the procedure is same as processing MP2MP-D Label Mapping message defined in [I-D.ietf-mpls-ldp-p2mp] section 4.3.1.5, and the processing protocol entity is HSMP-D label mapping message. The different procedure is specified below.

Node Z checks if upstream LSR U already assigned a label Lu to upstream <X, Y>. If not, transit node Z waits until it receives a HSMP-U Label Map <X, Y, Lu> from LSR U. Once the HSMP-U Label Map is received from LSR U, node Z checks whether it already has forwarding state upstream <X, Y> with incoming label Lu' and outgoing label Lu. If it does, Z sends a HSMP-U Label Map <X, Y, Lu'> to downstream node. If it does not, it allocates a label Lu' and creates a new label swap for Lu' with Label Lu over interface Iu. Interface Iu is determined via the procedures in Section 4.3.1. Node Z determines the downstream HSMP LSR as per Section 4.3.1, and sends a HSMP-U Label Map <X, Y, Lu'> to node D.

Since a packet from any downstream node is forwarded only to the upstream node, the same label (representing the upstream path) can be distributed to all downstream nodes. This differs from the procedures for MPMP LSPs [I-D.ietf-mpls-ldp-p2mp], where a distinct label must be distributed to each downstream node. The forwarding state upstream <X, Y> on node Z will be like this {<Lu'>, <Iu Lu>}. Iu means the upstream interface over which Z receives HSMP-U Label Map <X, Y, Lu> from LSR U. Packets from any downstream interface over which Z send HSMP-U Label Map <X, Y, Lu'> with label Lu' will be forwarded to Iu with label Lu' swap to Lu.

4.3.1.3. HSMP LSP root node operation

Suppose root node Z receives a HSMP-D Label Map <X, Y, L> from node D, the procedure is much same as processing MP2MP-D Label Mapping message defined in [I-D.ietf-mpls-ldp-p2mp] section 4.3.1.6, and the processing protocol entity is HSMP-D label mapping message. The different procedure is specified below.

Node Z checks if it has forwarding state for upstream <X, Y>. If not, Z creates a forwarding state for incoming label Lu' that indicates that Z is the LSP egress. E.g., the forwarding state might specify that the label stack is popped and the packet passed to some specific application. Node Z determines the downstream HSMP LSR as per section 4.3.1, and sends a HSMP-U Label Map <X, Y, Lu'> to node D.

Since Z is the root of the tree, Z will not send a HSMP-D Label Map and will not receive a HSMP-U Label Map.

4.3.2. HSMP LSP Label Withdraw

The HSMP Label Withdraw procedure is much same as MP2MP leaf operation defined in [I-D.ietf-mpls-ldp-p2mp] section 4.3.2, and the processing protocol entities are HSMP FECs. The only difference is process of HSMP-U label release message, which is specified below.

When a transit node Z receives a HSMP-U label release message from downstream node D, Z should check if there are any incoming interface in forwarding state upstream <X, Y>. If all downstream nodes are released and there is no incoming interface, Z should delete the forwarding state upstream <X, Y> and send HSMP-U label release message to its upstream node.

4.3.3. HSMP LSP upstream LSR change

The procedure for changing the upstream LSR is the same as defined in [I-D.ietf-mpls-ldp-p2mp] section 4.3.3, except it is applied to HSMP

FECs.

5. HSMP LSP on a LAN

The procedure to process P2MP LSP on a LAN has been described in [I-D.ietf-mpls-ldp-p2mp]. When the LSR forwards a packet downstream on one of those LSPs, the packet's top label must be the "upstream LSR label", and the packet's second label is "LSP label".

When establishing the downstream path of a HSMP LSP, as defined in [I-D.ietf-mpls-ldp-upstream], a label request for a LSP label is send to the upstream LSR. The upstream LSR should send label mapping that contains the LSP label for the downstream HSMP FEC and the upstream LSR context label. At the same time, it must also send label mapping for upstream HSMP FEC to downstream node. Packets sent by the upstream router can be forwarded downstream using this forwarding state based on a two label lookup. Packets traveling upstream need to be forwarded in the direction of the root by using the label allocated by upstream LSR.

6. Redundancy considerations

In some scenario, it is necessary to provide two root nodes for redundancy purpose. One way to implement this is to use two independent HSMP LSPs acting as active/standby. At one time, only one HSMP LSP will be active, and the other will be standby. After detecting the failure of active HSMP LSP, the root and leaf nodes will switch the traffic to the new active HSMP LSP which is switched from former standby LSP. The detail of redundancy mechanism will be for future study.

7. Security Considerations

The same security considerations apply as for the MP2MP LSP described in [I-D.ietf-mpls-ldp-p2mp].

8. IANA Considerations

This document requires allocation of two new LDP FEC Element types:

1. the HSMP-upstream FEC type - requested value 0x09

2. the HSMP-downstream FEC type - requested value 0x10

This document requires the assignment of new code points for the Capability Parameter TLVs, corresponding to the advertisement of the HSMP LSP capabilities. The values requested are:

HSMP LSP Capability Parameter - requested value 0x050B

9. Acknowledgement

The author would like to thank Eric Rosen, Sebastien Jobert, Fei Su, Edward for their valuable comments.

10. References

10.1. Normative references

[I-D.ietf-mpls-ldp-p2mp]
Minei, I., Wijnands, I., Kompella, K., and B. Thomas, "Label Distribution Protocol Extensions for Point-to-Multipoint and Multipoint-to-Multipoint Label Switched Paths", draft-ietf-mpls-ldp-p2mp-15 (work in progress), August 2011.

[I-D.ietf-mpls-ldp-upstream]
Aggarwal, R. and J. Roux, "MPLS Upstream Label Assignment for LDP", draft-ietf-mpls-ldp-upstream-10 (work in progress), February 2011.

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

[RFC5332] Eckert, T., Rosen, E., Aggarwal, R., and Y. Rekhter, "MPLS Multicast Encapsulations", RFC 5332, August 2008.

[RFC5561] Thomas, B., Raza, K., Aggarwal, S., Aggarwal, R., and JL. Le Roux, "LDP Capabilities", RFC 5561, July 2009.

10.2. Informative References

[I-D.ietf-l2vpn-vpms-frmwk-requirements]
Kamite, Y., JOUNAY, F., Niven-Jenkins, B., Brungard, D., and L. Jin, "Framework and Requirements for Virtual Private Multicast Service (VPMS)", draft-ietf-l2vpn-vpms-frmwk-requirements-04 (work in progress), July 2011.

[I-D.ietf-mpls-loss-delay]

Frost, D. and S. Bryant, "Packet Loss and Delay Measurement for MPLS Networks", draft-ietf-mpls-loss-delay-04 (work in progress), July 2011.

[I-D.ietf-pwe3-p2mp-pw]
Martini, L., Boutros, S., Sivabalan, S., Konstantynowicz, M., Vecchio, G., Nadeau, T., JOUNAY, F., Niger, P., Kamite, Y., Jin, L., Vigoureux, M., Ciavaglia, L., and S. Delord, "Signaling Root-Initiated Point-to-Multipoint Pseudowires using LDP", draft-ietf-pwe3-p2mp-pw-02 (work in progress), March 2011.

[I-D.ietf-tictoc-1588overmpls]
Davari, S., Oren, A., Bhatia, M., Roberts, P., and L. Montini, "Transporting PTP messages (1588) over MPLS Networks", draft-ietf-tictoc-1588overmpls-01 (work in progress), May 2011.

[IEEE1588v2]
"IEEE standard for a precision clock synchronization protocol for networked measurement and control systems", IEEE1588v2 , March 2008.

[RFC4762] Lasserre, M. and V. Kompella, "Virtual Private LAN Service (VPLS) Using Label Distribution Protocol (LDP) Signaling", RFC 4762, January 2007.

Authors' Addresses

Lizhong Jin
ZTE Corporation
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

IJsbrand Wijnands
Cisco Systems, Inc
De kleetlaan 6a
Diegem 1831, Belgium

Email: ice@cisco.com

Nicolai Leymann
Deutsche Telekom AG
Winterfeldtstrasse 21
Berlin 10781

Email: N.Leymann@telekom.de

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: May 3, 2012

L. Jin
ZTE
K. Liu
Nokia Siemens
S. Kini
Ericsson
October 31, 2011

Leaf discovery mechanism for mLDP based P2MP/MP2MP LSP
draft-jin-mpls-mldp-leaf-discovery-03.txt

Abstract

This document describes a mechanism for a root node to discover the leaf nodes of an mLDP based P2MP/MP2MP LSP. Such kind of function could be used for multiplexing/aggregating root initiated and leaf initiated application which will use mLDP based P2MP/MP2MP LSP. Examples of root initiated applications are P2MP PW [I-D.ietf-pwe3-p2mp-pw], VPLS multicast [I-D.ietf-l2vpn-vpls-mcast], L3VPN multicast [I-D.ietf-l3vpn-2547bis-mcast]. And examples of leaf initiated applications are statically configured mLDP based P2MP/MP2MP LSP, mLDP in-band signaling [I-D.ietf-mpls-mldp-in-band-signaling].

Status of this Memo

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

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

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

This Internet-Draft will expire on May 3, 2012.

Copyright Notice

Table of Contents

1. Introduction

This document describes a mechanism for a root node to discover the leaf nodes of an mLDP based P2MP/MP2MP LSP. Such kind of function could be used for multiplexing/aggregating root initiated and leaf initiated application which will use mLDP based P2MP/MP2MP LSP. Examples of root initiated applications are P2MP PW [I-D.ietf-pwe3-p2mp-pw], VPLS multicast [I-D.ietf-l2vpn-vpls-mcast], L3VPN multicast [I-D.ietf-l3vpn-2547bis-mcast]. And examples of leaf initiated applications are statically configured mLDP based P2MP/ MP2MP LSP, mLDP in-band signaling [I-D.ietf-mpls-mldp-in-band-signaling].

This draft provides a discovery mechanism based on a signaling session between each leaf and root node. Each leaf node would signal the leaf node information to root node through this session. There are two signaling protocols to be used for root initiated application, targeted LDP [RFC5036] or BGP auto-discovery using BGP Multiprotocol Extensions [RFC4760]. In order to reuse the signaling protocol of root initiated application, this document introduces both signaling protocols for mLDP leaf discovery.

2. Motivation and problem statement

The leaf initiated application mLDP in-band signaled P2MP LSP will trigger the leaf node to join from leaf node, which means none of the members belonging to a P2MP/MP2MP LSP topology knows all the other members of the P2MP/MP2MP LSP. This means that the root node cannot get the whole P2MP/MP2MP LSP membership information. This problem may cause some limitation for multiplexing/aggregation root initiated applications using mLDP LSPs.

Multicast VPLS [I-D.ietf-l2vpn-vpls-mcast] is a root initiated application. When setting up a inclusive P-Multicast tunnel, BGP A-D is used to do the VPLS membership auto-discovery. The mLDP based P2MP/MP2MP LSP will be set up when receiving auto-discovery routes through BGP A-D. The root node will only know the mLDP LSP leaf node information which is triggered by the specific BGP A-D mechanism. Let’s assume that a mLDP in-band signaling P2MP/MP2MP LSP_a (setup by leaf initiated application) already exist on the root node, and that LSP_a has the same leaf nodes as the P2MP LSP that VPLS multicast BGP A-D tries to set up. The root node does not know LSP_a leaf node information, and will set up mLDP based LSP_b triggered by BGP A-D with same root and leaf nodes.

This causes mLDP based LSP resources waste in the network as it may not be necessary to setup two mLDP LSPs with the same root and leaves

in the same network.

The introduction of a leaf discovery mechanism for mLDP based P2MP/ MP2MP LSP will enable leaf initiated applications to share one P2MP/ MP2MP LSP with root initiated application of P2MP/MP2MP LSP by multiplexing/aggregating mechanism.

3. Terminology

mLDP: Multicast LDP.

T-LDP: Target LDP.

P2MP LSP: An LSP that has one Ingress LSR and one or more Egress LSRs.

MP2MP LSP: An LSP that connects a set of nodes, such that traffic sent by any node in the LSP is delivered to all others.

Bud LSR: An LSR that is an egress but also has one or more directly connected downstream LSRs.

Ingress LSR: Source of the P2MP LSP, also referred to as root node.

Egress LSR: One of potentially many destinations of an LSP, also referred to as leaf node in the case of P2MP and MP2MP LSPs.

Transit LSR: An LSR that has one or more directly connected downstream LSRs.

Leaf node: A Leaf node can be either an Egress or Bud LSR when referred in the context of a P2MP LSP. In the context of a MP2MP LSP, an LSR is both Ingress and Egress for the same MP2MP LSP and can also be a Bud LSR.

P2MP FEC: The P2MP FEC Element consists of the address of the root of the P2MP LSP and an opaque value.

MP2MP FEC: MP2MP FEC consists of MP2MP downstream FEC and upstream FEC Element.

MP FEC: Includes both P2MP FEC and MP2MP FEC.

4. Leaf discovery mechanism

It would be beneficial if the mLDP leaf discovery mechanism can reuse

the same signaling session as the root initiated application, without requiring additional session overload. This document defines two leaf discovery mechanisms, one is based on T-LDP, the other is based on MP-BGP. Generally, the root initiated application with LDP as the main signaling mechanism, e.g, P2MP PW [I-D.ietf-pwe3-p2mp-pw], would use leaf discovery mechanism based on T-LDP, while application with MP-BGP as main signaling mechanism, e.g, VPLS Multicast[I-D.ietf-l2vpn-vpls-mcast], L3VPN Multicast [I-D.ietf-l3vpn-2547bis-mcast] may use leaf discovery mechanism based on MP-BGP.

4.1. Leaf discovery mechanism based on T-LDP

This section will introduce the discovery mechanism based on T-LDP session. Each leaf node will report the leaf node information to root through this T-LDP session. It is required that there is a T-LDP session existed between each leaf node and root node. mLDP leaf discovery function will share the same mLDP P2MP capability described in section 2.1 of [I-D.ietf-mpls-ldp-p2mp]

A LDP Label mapping message on the T-LDP session to the root with the MP FEC Element is used to convey the addition of the leaf membership to the root. The implicit NULL label is used to indicate that the mapping is from a leaf node. The Label Withdraw message is used to convey the deletion of the leaf membership to the root.

4.1.1. Node operation

The mLDP based P2MP/MP2MP LSP leaf discovery mechanism can be operated as follows.

For every leaf node, there will be a T-LDP session to be setup between root and leaf node. This T-LDP session can be setup automatically or manually, which depends on specific implementation.

When the leaf node is triggered to join one P2MP/MP2MP LSP, by various applications, the leaf node sends label mapping message to its upstream node (root or transit node). At the same time, the leaf node sends LDP label map message with MP FEC to its root node. When the root node receives the LDP label map message over T-LDP session with MP FEC, it will store the leaf node information associated with the specified P2MP/MP2MP LSP locally.

When the leaf node is triggered to leave one P2MP/MP2MP LSP, by various applications, the leaf node sends label withdraw message to its upstream node (root or transit node). At the same time, the leaf node sends LDP label withdraw message with MP FEC to its root node. When the root node receives the LDP label withdraw message over T-LDP

with MP FEC, it will delete the leaf node information associated with the specified P2MP/MP2MP LSP locally.

4.2. Leaf discovery mechanism based on MP-BGP

This section will introduce the discovery mechanism based on MP-BGP[RFC4760]. Each leaf node will report the leaf node information to root through this BGP session.

4.2.1. mLDP leaf NLRI

This document defines a new BGP NLRI, called mLDP leaf NLRI. Following is the format of the mLDP leaf NLRI:

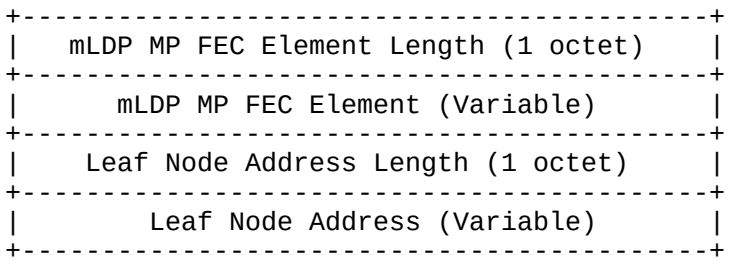


Figure 1

mLDP MP FEC Element may either contain P2MP FEC or MP2MP FEC element. Leaf Node Address field contains the leaf node IP address, and the value of length is 32 if it is IPv4 address, or 128 if it is IPv6 address. The NLRI field in the MP_REACH_NLRI and MP_UNREACH_NLRI is a mLDP MP FEC Element attached with Leaf Node Address. The mLDP leaf NLRI is advertised in BGP UPDATE messages using the MP_REACH_NLRI and MP_UNREACH_NLRI attributes [RFC4760]. The [AFI, SAFI] value pair used to identify this NLRI is (AFI=26(AFI for MPLS Multicast, pending, IANA allocation), SAFI=8(SAFI for mLDP leaf discovery, pending IANA allocation)).

In order for two BGP speakers to exchange mLDP leaf NLRI, they must use BGP Capabilities Advertisement to ensure that they both are capable of properly processing such NLRI. This is done as specified in [RFC4760], by using capability code 1 (multiprotocol BGP) with an AFI of 26 and an SAFI of mLDP leaf discovery.

The Next Hop field of MP_REACH_NLRI attribute shall be interpreted as an IPv4 address, whenever the length of the NextHop address is 4 octets, and as a IPv6 address, whenever the length of the NextHop address is 16 octets.

4.2.2. Node operation

The mLDP based P2MP/MP2MP LSP leaf discovery mechanism can be operated as follows.

When the leaf node is triggered to join one P2MP/MP2MP LSP, by various applications, the leaf node sends label mapping message to its upstream node (root or transit node). At the same time, the leaf node sends BGP UPDATE messages with MP_REACH_NLRI to its root node. The mLDP leaf NLRI will set Leaf Node Address to leaf node IP address, and next hop field to leaf node identifier. When the root node receives BGP UPDATE messages with MP_REACH_NLRI, it will store the leaf node information associated with the specified P2MP/MP2MP LSP locally.

When the leaf node is triggered to leave one P2MP/MP2MP LSP, by various applications, the leaf node sends label withdraw message to its upstream node (root or transit node). At the same time, the leaf node sends BGP UPDATE messages with MP_UNREACH_NLRI to its root node. The mLDP leaf NLRI will set Leaf Node Address to leaf node IP address, and next hop field to leaf node identifier. When the root node receives BGP UPDATE messages with MP_UNREACH_NLRI, it will delete the leaf node information associated with the specified P2MP/MP2MP LSP locally.

To constrain distribution of the mLDP leaf NLRI to the AS of the advertising PE the BGP Update message originated by the advertising PE SHOULD carry the NO_EXPORT Community [RFC1997].

5. Scalability

As recommended in section 4, leaf discovery will reuse the same signaling session as application, and will not setup additional sessions. For the application that uses T-LDP to do leaf discovery, all the leaf nodes have to setup T-LDP session to root node. There may be too many T-LDP sessions that have to be setup on the root node in the network, which will cause some scalability problem. This problem is caused by the application and out of scope of this draft.

6. Security Considerations

The same security considerations apply as for the multicast LDP specification, as described in [I-D. draft-ietf-mpls-ldp-p2mp], and MP-BGP, as described in [RFC4760].

7. IANA Considerations

7.1. MP-BGP

This document requires allocation of a new BGP AFI and SAFI.

A new AFI is allocated for MPLS Multicast function, the requested value has been pre-allocated as 26.

A new BGP SAFI for "Network Layer Reachability Information used for mLDP leaf discovery" from the IANA "Subsequence Address Family Identifiers (SAFI)" registry. The requested value has been pre-allocated as 8.

8. Acknowledgement

The author would like to thank Rahul Aggarwal, Dimitri Papadimitriou, IJsbrand Wijnands, Sandeep Bishnoi, Frederic Jounay and Simon DeLord for their valuable comments.

9. References

9.1. Normative references

[RFC1997] Chandrasekeran, R., Traina, P., and T. Li, "BGP Communities Attribute", RFC 1997, August 1996.

[RFC4760] Bates, T., Chandra, R., Katz, D., and Y. Rekhter, "Multiprotocol Extensions for BGP-4", RFC 4760, January 2007.

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

9.2. Informative References

[I-D.ietf-l2vpn-vpls-mcast] Aggarwal, R., Kamite, Y., and L. Fang, "Multicast in VPLS", draft-ietf-l2vpn-vpls-mcast-09 (work in progress), July 2011.

[I-D.ietf-l3vpn-2547bis-mcast] Aggarwal, R., Bandi, S., Cai, Y., Morin, T., Rekhter, Y., Rosen, E., Wijnands, I., and S. Yasukawa, "Multicast in MPLS/BGP IP VPNs", draft-ietf-l3vpn-2547bis-mcast-10 (work in progress), January 2010.

[I-D.ietf-mpls-ldp-p2mp]
Minei, I., Wijnands, I., Kompella, K., and B. Thomas, "Label Distribution Protocol Extensions for Point-to-Multipoint and Multipoint-to-Multipoint Label Switched Paths", draft-ietf-mpls-ldp-p2mp-15 (work in progress), August 2011.

[I-D.ietf-mpls-mldp-in-band-signaling]
Wijnands, I., Eckert, T., Leymann, N., and M. Napierala, "Multipoint LDP in-band signaling for Point-to-Multipoint and Multipoint- to-Multipoint Label Switched Paths", draft-ietf-mpls-mldp-in-band-signaling-04 (work in progress), May 2011.

[I-D.ietf-pwe3-p2mp-pw]
Sivabalan, S., Boutros, S., Martini, L., Konstantynowicz, M., Vecchio, G., Kamite, Y., and L. Jin, "Signaling Root-Initiated Point-to-Multipoint Pseudowire using LDP", draft-ietf-pwe3-p2mp-pw-03 (work in progress), October 2011.

[I-D.ietf-pwe3-p2mp-pw-requirements]
Heron, G., Wang, L., Aggarwal, R., Vigoureux, M., Bocci, M., Jin, L., JOUNAY, F., Niger, P., Kamite, Y., DeLord, S., and L. Martini, "Requirements and Framework for Point-to-Multipoint Pseudowires over MPLS PSNs", draft-ietf-pwe3-p2mp-pw-requirements-05 (work in progress), September 2011.

Authors' Addresses

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

Email: lizhong.jin@zte.com.cn

Kebo Liu
Nokia Siemens Networks
1122 North Qinzhou Road
Shanghai, 200233, China

Email: kebo.liu@nsn.com

Sriganesh Kini
Ericsson
300 Holger Way
San Jose, CA 95134

Email: sriganesh.kini@ericsson.com

MPLS Working Group S. Kini
Internet Draft S. Narayanan
Intended Status: Standards Track Ericsson
Expires: May 2012 October 31, 2011

MPLS Fast Re-route using extensions to LDP
draft-kini-mpls-frr-ldp-02.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.

Table of Contents

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) for a shortest path LSP (or FEC) is another LSP that goes from the PLR to an LSR that can label-switch the traffic back along that part of the shortest-path LSP that is not upstream for that failure. The LSR on which the BSP LSP terminates is called the BSP Merge Point or BSP-MP.

The BSP LSPs are LDP 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] is used. The mechanisms in this draft should be used when an LFA is not available.

A shortest path LSP to the BSP-MP should be used as a BSP LSP if one is available. This must use label stacking as follows. First the label of the LSP should be swapped with that allocated by the BSP-MP for that FEC. Next, the label to the BSP-MP should be stacked. When a single such shortest path LSP is not available to be used as a BSP LSP, multiple shortest path LSPs and/or interfaces on directly connected LSRs can be stitched together. The LSRs that stitch such LSPs together do so by advertising another label for the FEC. This label is stacked below the shortest path LSP label. It is allocated on demand and is initiated from the PLR to the first stitching LSR. If there is a stitching LSR further downstream (i.e. towards the BSP-MP) the stitching LSR in turn requests a label from the downstream stitching LSR.

The protocol extensions required to setup BSP LSPs are described in section 4. The label actions for the BSP LSPs are pre-installed in the forwarding tables. The PLR pre-computes the label-operation changes to be performed on the failure trigger. When the failure occurs, the PLR detects the failure as a local event and performs the pre-computed label operation actions. None of the LSRs along the BSP LSP other than the PLR have to perform any additional operation at the instant of failure in order to protect the traffic. FRR-LDP is a local repair mechanism that can protect against link, node and SRLG failures.

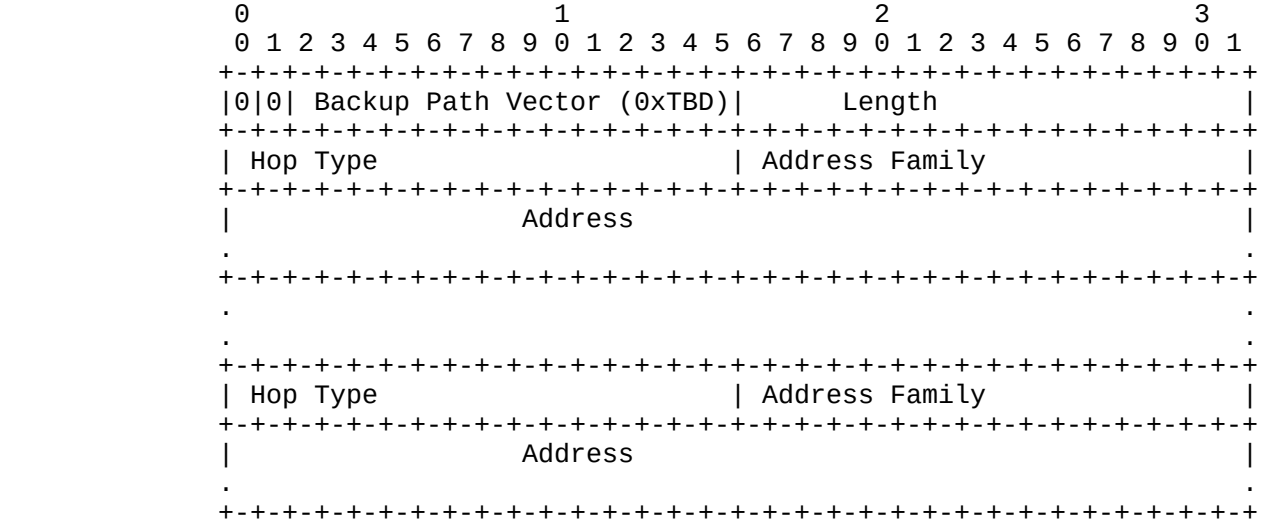
The BSP LSPs that have to be setup depend on the network topology. It also depends on the type of protection (e.g. per-FEC or per-nexthop). Sample topologies with FRR-LDP applied are described in section 5.

5. Protocol procedures and extensions

The PLR must establish a targeted LDP session with the BSP MP or the next stitching LSR to get the label bindings it needs for the backup LSP. It must negotiate the on-demand mode for the targeted sessions

and request only those label bindings that it needs to protect the LSPs. Additionally the PLR sends the path information when the next LSR is a stitching LSR so that it can request the stitching LSR further downstream for label bindings. A new TLV "Backup Path Vector TLV" is defined. It should be noted that this TLV is required only when multiple shortest path LSPs need to be stitched together for the BSP LSP.

The TLV consists of a list of addresses of stitching LSRs.



Hop Type

1 - Indicates that the stitching LSR should use a shortest path LDP LSP to reach the next stitching LSR

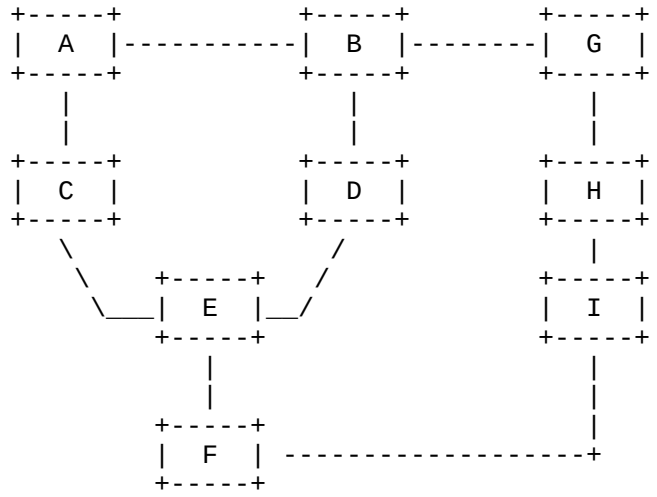
2 - Indicates that the stitching LSR should use a directly connected interface to reach the next stitching LSR

This TLV is used in the "Label Request" message. Each stitching LSR removes the first address in this TLV before requesting a label from the next stitching LSR.

A Capability TLV is defined for LDP in accordance to [LDP-CAP] to advertise its capability to process the "Backup Path Vector" TLV. An area scope capability TLV is also advertised via IGP ([OSPF-CAP] and [ISIS-CAP]) for the same so that the PLR can take it into account when computing the BSP LSP.

6. Examples of FRR-LDP

In the example topology below, link A-B is of cost 100. All other links are unit cost. To protect against the failure of LSR E for FEC F, LSR C sets up a BSP LSP C-A-B-G where A and B stitch shortest path LSP C-A, the LSP A-B (due to B advertising a label to A) and the shortest path LSP B-G. The PLR (LSR C) computes the backup path by executing a shortest path computation on the topology excluding LSR E. It initiates a label-request towards LSR A for FEC F, with the "Backup Path Vector" TLV containing the address of LSR B (Hop Type 2) and of LSR G (Hop Type 1). LSR A further initiates a label-request towards LSR B with a "Backup Path Vector" TLV containing the address of LSR G (Hop Type 1). LSR B in turn initiates a label request for FEC F to LSR G (without the Backup Path Vector TLV). Note that in this case since B and G are directly connected this label may be the same one that LSR G had originally allocated for F. LSR B then allocates a label F, sets the label operation to swap to the label allocated by G and to forward over LSP B-G and returns that label in the label response to LSR A. LSR A in turn allocates another label for F, sets the label operation to swap to the label allocated by B and to forward over LSP A-B and returns that label in the label response to LSR C.



7. Security Considerations

This document does not introduce any additional security considerations beyond those in [LDP].

8. IANA Considerations

New TLV types are needed for the "Backup Path Vector" and the LDP,

OSPF and ISIS Capability values.

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.

[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.

[OSPF-CAP] Lindem, A., et al, "Extensions to OSPF for Advertising Optional Router Capabilities", RFC 4970, July 2007.

[ISIS-CAP] Vasseur, JP., et al, "Intermediate System to Intermediate System (IS-IS) Extensions for Advertising Router Information", RFC 4971, July 2007.

[IPFRR-LFA] Atlas, A., et al, "Basic Specification for IP Fast Reroute: Loop-Free Alternates", RFC 5286, September 2008.

[LDP-CAP] Thomas, B., et al, "LDP Capabilities", RFC 5561, July 2009.

9.2. 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.

10. 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: April 30, 2012 October 31, 2011

Temporal and hitless path segment monitoring
draft-koike-mpls-tp-temporal-hitless-psm-04.txt

Status of this Memo

Copyright Notice

Abstract

The MPLS transport profile (MPLS-TP) is being standardized to enable carrier-grade packet transport and complement converged packet network deployments. Among 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 is 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 Sub-path Maintenance Elements (SPMEs) which provides monitoring of a portion of a set of transport paths (LSPs or MS-PWs). Based on the problems, this document specifies new requirements to consider a new improved mechanism of hitless transport path 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

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 per se 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.

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 and proposes to consider a new improved method of the segment monitoring, following up the work done in [5]. Moreover, this document explains detailed requirements on the new temporal and hitless segment monitoring function which are not covered in [5].

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

HSPM Hitless Path Segment Monitoring

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 for MPLS-TP networks 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 is common in transport networks that network objective (1) is mandatory and that regarding network objective (2) the monitoring shall not change the forwarding behavior.

4. Problem statement

To monitor, protect, or manage portions of transport paths, such as LSPs 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 transport path 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 case of 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 only for maintenance purposes. As the size of monitored segments increases, the size of the label stack grows. Moreover, if the 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 only the monitoring purpose should be avoided to maximize the available bandwidth.

Problem (P-2) is related to an identifier-management issue. The identification of each layer in case of LSP label stacking is required in terms of strict sub-layer management for the segment monitoring in a MPLS-TP network. There is no standardized way to identify a layer, however a possible rule of differentiating layers will be necessary at least within an administrative domain, if SPME is applied for on-demand OAM functions. This enforces operators to create an additional permanent layer identification policy only for temporal path segment monitoring. 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 identifiers and managed layers 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). Moreover, as a result of these pre-configurations, they impose operators to pre-design the structure of sub-path maintenance elements, which is not preferable in terms of operators’ increased burden. These concerns are summarized in section 3.8 of [5].

Furthermore, in reality, all the possible patterns of path segment cannot be set in SPME, because overlapping of path segments is limited to nesting relationship. As a result, possible SPME patterns of portions of an original transport path are limited due to the characteristic of SPME shown in Figure.1, even if SPMEs are pre-configured. This restriction is inconvenient when operators have to fix issues in an on-demand manner.To avoid these issues, the temporal and on-demand setting of the SPME(s) is needed and more efficient for monitoring in MPLS-TP transport network operation.

However, using currently defined methods, the temporal setting of SPMEs also causes 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 all the MPLS frames and changing label value(s)

(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. As shown in network objective (2), the monitoring function needs to monitor 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.

Another aspect is that 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 differs from the original data plane conditions.

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. As shown in Fig.1, SPME changes the length of all the MPLS frames and changes label value(s). This is no longer the monitoring of the original transport path but the monitoring of a different path. Particularly, performance monitoring measurement (Delay measurement and loss measurement) are sensitive to those changes.

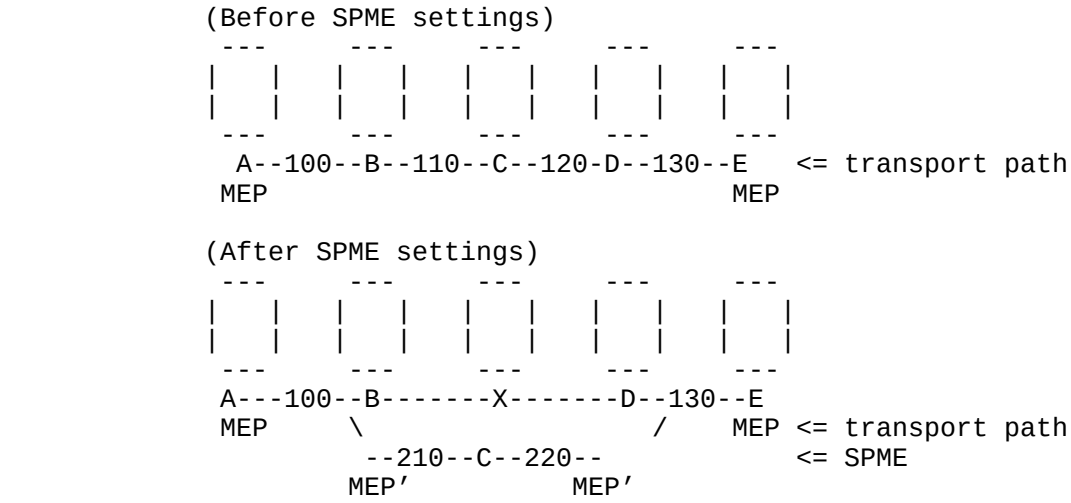


Figure 1 : An Example of a SPME setting

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 conditions of the targeted transport path, because the make-before-break procedure is premised on the change of the inner label value. This means changing one of the settings in MPLS shim header.

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 traffic) as the creation of an SPME on the same LSP.

Note: (P'-2) will be removed when non-disruptive make-before-break (in both with and without C-plane environment) is specified in other MPLS-TP documents. However, (P'-2) could be replaced with the following issue. ''Non-disruptive MBB, in other words, taking an action similar to switching just for monitoring is not an ideal operation in transport networks.

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 could not be found 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 utility of SPMEs is basically limited to inter-carrier or inter-domain segment monitoring where they are typically pre-configured or pre-instantiated. SPME 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. SPME is expected to be mainly used for protection purpose within one administrative domain.

To summarize, the problem statement is that the current sub-path maintenance based on a hierarchical LSP (SPME) is problematic for pre-configuration in terms of increasing bandwidth by label stacking and managing objects by layer stacking and address management. A on-demand/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. Another monitoring mechanism is therefore required that supports temporal and hitless path segment monitoring. Hereafter it is called on-demand hitless path segment monitoring (HPSM).

Note: The above sentence ''and disrupt the in-service customer traffic'' might need to be modified depending on the result of future discussion about (P'-2).

5. OAM functions using segment monitoring

OAM functions in which on-demand HPSM 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. In other words, the characteristic of "on-demand" is generally temporal for maintenance operation. Conversely, this could be a good reason that operations should not be based on pre-configuration and pre-design.

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, on-demand HPSM for packet loss and packet delay measurements are indispensable for transport network operation.

Regarding other on-demand monitoring functions path 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 HPSM. However, specific segment monitoring should be applied to the OAM function of diagnostic test, because SPME doesn’t meet network objective (2) in section 3. See section 6.3.

Note:

The solution for temporal and hitless segment monitoring should not be limited to label stacking mechanisms based on pre-configuration, such as PST/TCM(label stacking), which can cause the issues (P-1) and (P-2) described in Section 4.

The solution for HPSM 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

6.1. Necessity of on-demand single-level monitoring

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-level 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-level and simultaneous monitoring is the most powerful tool for accurately diagnosing the performance of a transport path. However, considering the substantial benefits to operators, a strict monitoring function which is required in such as a test environment of a laboratory does not seem to be necessary in the field. To summarize, on-demand and in-service (hitless) single-level segment monitoring is required, on-demand and in-service multi-level segment monitoring isdesirable. Figure 2 shows an example of a multi-level on-demand segment monitoring.

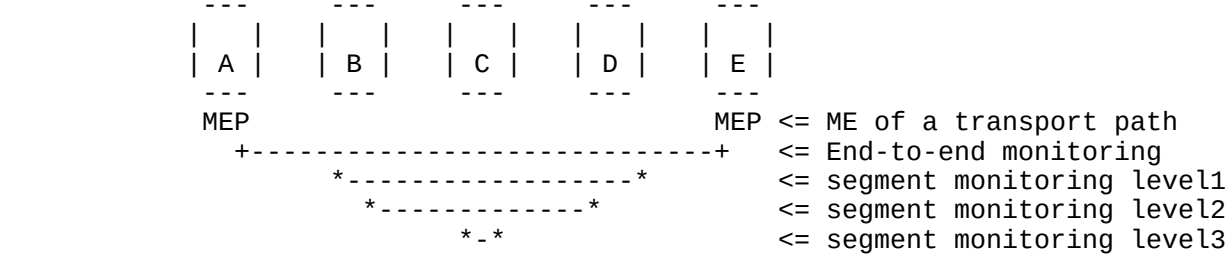


Figure 2 : An Example of a multi-level on-demand segment monitoring

6.2. Necessity of on-demand monitoring independent from end-to-end proactive monitoring

As multi-level simultaneous monitoring only for on-demand new path segment monitoring was already discussed in section6.1, next we consider the necessity of simultaneous monitoring of end-to-end current proactive monitoring and new on-demand path segment monitoring. Normally, the on-demand path segment monitoring is configured in a segment of a maintenance entity of a transport path. In this environment, on-demand single-level monitoring should be done without disrupting pro-active monitoring of the targeted end-to-end transport path.

If operators have to disable the pro-active monitoring during the on-demand hitless path 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.

Accordingly, the on-demand single lavel path segment monitoring is required without changing or interfering the proactive monitoring of the original end-to-end transport path.

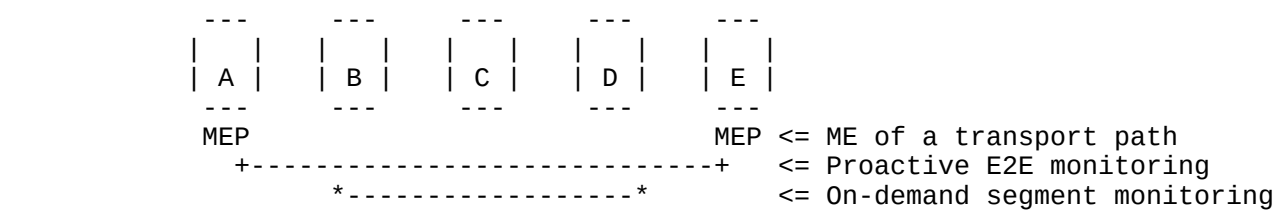


Figure 3 : Independency between proactive end-to-end monitoring and on-demand segment monitoring

6.3. Necessity of arbitrary segment monitoring

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).

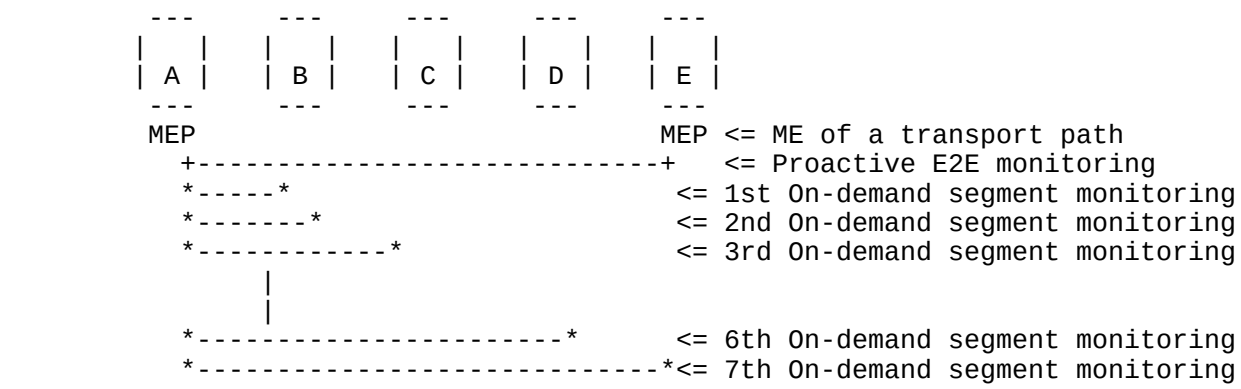


Figure 4 : One possible procedure to localize a fault point by sequential on-demand segment monitoring

Accordingly, on-demand monitoring of arbitrary segments is mandatory in the case in Fig. 5. As a result, on-demand HSPM 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.

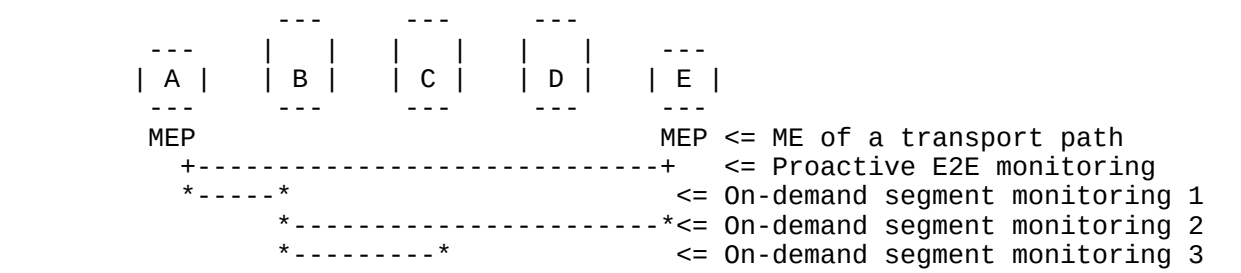


Figure 5 : Example where on-demand monitoring has to be configured in arbitrary segments

7. Conclusion

It is requested that another monitoring mechanism is required to support temporal and hitless segment monitoring which meets the two network objectives mentioned in Section 3 of this draftthat are described also 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.

The solution for the temporal and hitless segment monitoring has to cover both per-node model and per-interface model which are specified in [5]. In addition, the following requirements should be considered for an enhanced temporal and hitless path segment monitoring function.

Note: (P'-2) needs to be reconsidered.- An on-demand and in-service ''single-level'' segment monitoring ismandatory. Multi-level segment monitoring is optional.

- ''On-demand and in-service'' single level 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.

The followings are specific requirements on each on-demand OAM function.Mandatory: Packet Loss Measurement and Packet Delay Measurement

Option: Connectivity verification, Diagnostic Tests (Throughput test), Route tracing

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
Updates: 3209, 3473 (if approved)
Intended status: Standards Track
Expires: May 3, 2012

K. Kompella
Juniper Networks

October 31, 2011

Multi-path Label Switched Paths Signaled Using RSVP-TE
draft-kompella-mpls-rsvp-ecmp-01.txt

Abstract

This document describes extensions to Resource ReSerVation Protocol - Traffic Engineering for the set up of multi-path Traffic Engineered Label Switched Paths (LSPs) in Multi Protocol Label Switching and Generalized MPLS networks, i.e., LSPs that conform to traffic engineering constraints, but follow multiple independent paths from the source to the destination that allow load balancing.

Status of this Memo

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

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

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

This Internet-Draft will expire on May 3, 2012.

Copyright Notice

Table of Contents

1. Introduction

In selecting a protocol for setting up and signaling "tunnel" Labeled Switched Paths (LSPs) in Multi Protocol Label Switching (MPLS) and Generalized MPLS (GMPLS) networks, one first chooses whether one wants Equal Cost Multi-Path (ECMP) load balancing or Traffic Engineering (TE). For the former, one uses the Label Distribution Protocol (LDP) ([RFC5036]); for the latter, the Resource ReSerVation Protocol - Traffic Engineering (RSVP-TE) ([RFC3209]). [Two other criteria, the need for fast protection and the desire for less configuration, are no longer the deciding factors they used to be, thanks to "IP fast reroute" ([RFC5286]) and "RSVP-TE automesh" ([RFC4972])].

This document describes how one can set up a tunnel LSP that has both ECMP and TE characteristics using RSVP-TE. The techniques described in this document can be used to create a single overall "ECMP TE LSP" to a single destination that consists of several "sub-LSPs", each taking a different path through the network to the same destination. The techniques can also be used to create a single ECMP TE LSP to multiple equivalent destinations (such as equidistant BGP nexthops announcing a common set of reachable addresses), such that each destination is served by one or more sub-LSPs. The techniques described here borrow the notion of sub-LSPs from [RFC4875].

Several options are available for ECMP TE LSPs. One is that the ingress Label Switching Router (LSR) computes (or otherwise obtains) all sub-LSP paths; alternatively, LSRs along the various paths can compute paths further downstream (using techniques such as "loose hop expansion", as in [RFC5152]). Another is that an RSVP Path message can contain information about exactly one path through the network (or sub-LSP); alternately, a Path message can contain information about more than one such path. A third option is that the various paths that make up the multi-path LSP have equal cost (or distance) from ingress to egress (i.e., ECMP), as opposed to paths that may have differing costs. Another option (mentioned above) is to terminate a multi-path LSP on a single egress or on several equivalent egresses. For now, the first of each of these alternatives is assumed; future work can explore other choices.

1.1. Terminology

The terms "tunnel", "tunnel LSP" and "LSP" all refer to a container LSP from an ingress LSR to egress LSR(s). An LSP is the unit of configuration, signaling and management.

An ECMP (or generally, a multi-path) TE LSP is called a Multi-path Label Switched Path (MLSP), and consists of one or more sub-LSPs.

A sub-LSP consists of a single path from the ingress to one egress. A "regular" point-to-point TE LSP is equivalent to an MLSP with a single sub-LSP.

The "downstream links" of an LSR X with respect to an MLSP Z is the set of all links adjacent to X traversed after X by at least one sub-LSP of MLSP Z.

1.2. 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. Theory of Operation

2.1. Multi-path Label Switched Paths

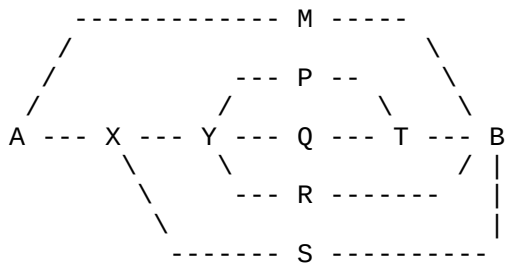
An MLSP is configured at the ingress with various constraints typically associated with TE LSPs, such as destination LSR(s), bandwidth (on a per-class basis, if desired), link colors, Shared Risk Link Groups, etc. [Auto-mesh techniques ([RFC4972]) can be used to reduce configuration; this is not described further here.] In addition, parameters specifically related to MLSPs, such as how many (or the maximum number of) sub-LSPs to create, whether traffic should be split equally across sub-LSPs or not, etc. may also be specified.

The ingress LSR can use the configuration parameters to decide how many sub-LSPs to compute for this MLSP and what paths they should take. Each sub-LSP MUST meet all the constraints of the MLSP (except the bandwidth). The bandwidths (per-class, if applicable) of all the sub-LSPs MUST add up to the bandwidth of the MLSP. If a Path Computation Element (PCE; [RFC4655]) that is multi-path LSP-aware is used, the PCE is subject to these same requirements; how MLSP requirements are signaled to a PCE is beyond the scope of this document.

Having computed (or otherwise obtained) the paths of all the sub-LSPs, the ingress A then signals the MLSP by signaling all the individual sub-LSPs across the MPLS/GMPLS network. If multiple sub-LSPs of the same MLSP pass through LSR Y, and Y has downstream links YP, YQ and YR for the various sub-LSPs, then Y has to load balance incoming traffic for the MLSP across the three downstream links in proportion to the sum of the bandwidths of the sub-LSPs going to each downstream (see Figure 1).

One must distinguish carefully between the (signaled) bandwidth of a sub-LSP, a static value capturing the expected or maximum traffic on the sub-LSP, and the instantaneous traffic received on a sub-LSP, a constantly varying quantity. Suppose there are three sub-LSPs traversing Y, with bandwidths 10Gbps, 20Gbps and 30Gbps, going to P, Q and R respectively. Suppose further Y receives some traffic over each of these sub-LSPs. Y must balance this received traffic over the three downstream links YP, YQ and YR in the ratio 1:2:3.

2.2. ECMP



An example network illustrating ECMP. Assume that paths AMB, AXYPTB, AXYQTB, AXYRB and AXSB all have the same path length (cost).

Figure 1: Example Network Topology

In an IP or LDP network, incoming traffic arriving at A headed for B will be split equally between M and X at A. Similarly, traffic for B arriving at Y will be split equally among P, Q and R. If the traffic arriving at A for B is 120Gbps, then the AMB path will carry 60Gbps, the paths AXYPTB, AXYQTB and AXYRB will each carry 10Gbps, and the AXSB path will carry 30Gbps. We'll call this "IP-style" load balancing.

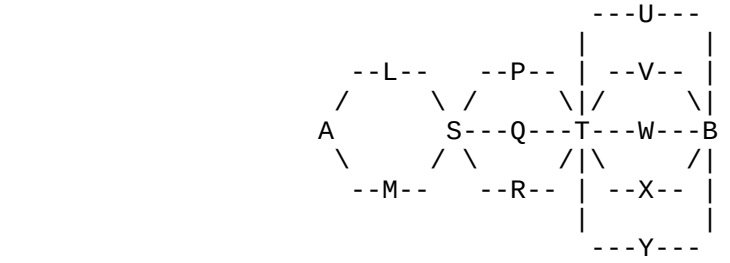
Note: all load balancing is subject to the overriding requirement of mapping the same "flow" to the same downstream. (What constitutes a "flow" is beyond the scope of this document.) This requirement takes precedence over all attempts to balance traffic among downstreams. Thus, the statements above (e.g., "the AMB path will carry 60Gbps") are to be interpreted as ideal targets, not hard requirements, of load balancing.

One can simulate the IP or LDP ECMP behavior with TE-based ECMP by creating an MLSP with five sub-LSPs S1 through S5 taking paths AMB, AXYPTB, AXYQTB, AXYRB and AXSB, with bandwidths 60Gbps, 10Gbps, 10Gbps, 10Gbps and 30Gbps, respectively.

With such an arrangement, the MB link carries 60Gbps while the RB link carries just 10Gbps. If one wishes instead to carry equal amounts of traffic on the links incoming to B, then one could arrange the sub-LSPs S1 to S5 to have bandwidths 30Gbps, 15Gbps, 15Gbps, 30Gbps and 30Gbps, respectively. In this case, the bandwidth on each of the four links going to B is 30Gbps, illustrating some of the capabilities of TE-based ECMP.

Staying with this example, A has one sub-LSP of bandwidth 30Gbps to M and four sub-LSPs of total bandwidth 90Gbps to X. Thus, A should load

balance traffic in the ratio 1:3 between the AM and the AX links. Similarly, X has three sub-LSPs of total bandwidth 60Gbps to Y and one sub-LSP of bandwidth 30Gbps to S, so X should load balance traffic 2:1 between Y and S. Y has a sub-LSP of bandwidth 15Gbps to each of P and Q and one sub-LSP of bandwidth 30Gbps to R, so Y should load balance traffic 1:1:2 among P, Q and R, respectively. Thus, in general, TE-based ECMP does not assume equal distribution of traffic among downstream LSRs, unlike IP- or LDP-style ECMP.



Another example network illustrating 30 ECMP paths between A and B.

Figure 2: Another Network Topology

In Figure 2, there are potentially 2x3x5=30 ECMP paths between A and B. With IP or LDP, exploiting all these paths is straightforward, and doesn't need a lot of state. With an MLSP as seen so far, this would require 30 sub-LSPs to achieve equivalent load balancing. This suggests that a different approach is needed to efficiently achieve IP-style load balancing with TE LSPs. To this end, we introduce the notion of "equi-bandwidth" (EB) sub-LSPs and EB MLSPs. A sub-LSP is equi-bandwidth if its "E" bit is set (see Section 3.1). An MLSP is equi-bandwidth if all of its sub-LSPs are equi-bandwidth.

If a set of EB sub-LSPs of the same MLSP traverse an LSR S, say to downstream links SP, SQ and SR, then S MUST attempt to load balance traffic received on these EB sub-LSPs equally among the links SP, SQ and SR, independent of how many sub-LSPs go over each of these links. Furthermore, S MUST redistribute traffic received from each of its upstream LSRs, and SHOULD redistribute all traffic received from upstream as a whole. One can do the former by signaling the same label to each of its upstream LSRs; one can do the latter by signaling the same label to all upstream LSRs (see Section 3.2). For example, in Figure 2, if L sends 12Gbps of traffic to S and M sends 18Gbps to S, S can redistribute L's traffic by sending 4Gbps to each of P, Q and R; and can similarly send 6Gbps of M's traffic to each of P, Q and R. Alternatively, S can load balance the aggregate 30Gbps of traffic received from L and M to each of P, Q and R, thus sending 10Gbps to each. EB sub-LSPs have an added benefit of not requiring

unequal load balancing across links, which may pose problems for some hardware.

Given the notion of EB sub-LSPs and EB MLSPs, A can signal an EB MLSP Z comprised of five EB sub-LSPs E1 through E5 with the following paths: ALSPTUB, AMSQTVB, ALSRTWB, AMSPTXB and ALSQTYB (respectively). Then, A has two downstream links for the five sub-LSPs, AL and AM, between which A will load balance equally. Similarly, S has three downstream links, SP, SQ and SR; and T has five downstreams, TU, TV, TW, TX and TY. Thus the load balancing behavior of the MLSP will replicate IP load balancing. The state required for an EB MLSP to achieve IP-style load balancing is somewhat greater than for LDP LSPs, but significantly less than that for multiple "regular" TE LSPs, or for a non-EB MLSP.

2.3. Discussion

Some of the power of TE-based ECMP was illustrated in the above examples. Another is ability to request that all sub-LSPs avoid links colored red. If in the example network in Figure 1, the QT link is colored red but all other links are not, then there are four ECMP paths that satisfy these constraints, and the traffic distribution among them will naturally be different than it would without the link color constraint.

One can also ask whether an MLSP with sub-LSPs is any better than N "regular" LSPs from the same ingress to the same egress. Here are some benefits of an MLSP:

1. With an MLSP, there is a single entity to provision, manage and monitor, versus N separate entities in the case of LSPs. A consequence of this is that with an MLSP, changes in topology can be dealt with easily and autonomously by the ingress LSR, by adding, changing or removing sub-LSPs to rebalance traffic, while maintaining the same TE constraints. With individual LSPs, such changes would require changes in configuration, and thus are harder to automate.

2. An ingress LSR, knowing that an MLSP is for load balancing, can decide on an optimum number of sub-LSPs, and place them appropriately across the network to optimize load balancing. On the other hand, an ingress LSR asked to create N independent LSPs will do so without regard to whether N is a good number of equal cost paths, and, more importantly, may place several of the N LSPs on the same path, defeating the purpose of load balancing.

3. The EB sub-LSP mechanism will, in many cases, result in far fewer sub-LSPs than independent LSPs and thus less control plane state.

4. Finally, an MLSP will usually have less data plane state than N independent LSPs: whenever multiple sub-LSPs traverse a link, a single label will be used for all of them, whereas if multiple LSPs traverse a link, each will need a separate label.

2.4. The Capabilities of TE-based Load Balancing

Definition: Let G=(V, E) be a directed graph (or network), and let A and B in V be two nodes in G. Let T be the traffic arriving at A destined for B. T is said to be "IP-style" load balanced if for every node X on a shortest path from A to B, the portion of T arriving at X is split equally among all nodes Yi that are adjacent to X and are on a shortest path from X to B.

Theorem: An MLSP can accurately mimic IP-style load balancing between any two nodes in any network.

Proof: left to the reader.

Corollary: MLSPs provide a strictly more powerful load balancing mechanism than IP-style load balancing.

3. Operation of MLSPs

3.1. Signaling MLSPs

An MLSP is identified by an LSP_TUNNEL SESSION object defined in [RFC3209]. All sub-LSPs of an MLSP have the same field values in their LSP_TUNNEL SESSION object.

A sub-LSP of an MLSP is identified by the LSP_TUNNEL SESSION object plus a new Sender Template object called the MLSP_TUNNEL Sender Template. The MLSP_TUNNEL Sender Template comes in two flavors, IPv4 and IPv6, shown below. The 15-bit Sub-LSP ID uniquely identifies a sub-LSP of an MLSP, and stays the same during the lifetime of the sub-LSP. The LSP ID may change as in [RFC3209] to let a sub-LSP share resources with another incarnation of the sub-LSP, for example to reroute and/or change bandwidths of the sub-LSP. The "E" bit defines whether a sub-LSP is an equi-bandwidth sub-LSP (E=1) or not (E=0). The equi-bandwidth character of a sub-LSP (i.e., the value of the E bit) MUST remain the same from ingress to egress as well as during the lifetime of a sub-LSP.

3.1.1. MLSP_TUNNEL Sender Template

Class = SENDER_TEMPLATE, MLSP_TUNNEL_IPv4 C-Type = TBD

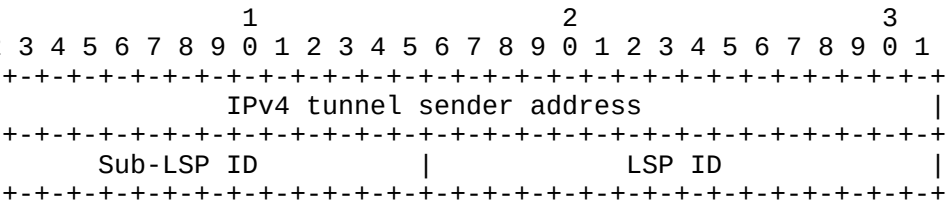


Figure 3: MLSP_TUNNEL_IPv4 Sender Template

Class = SENDER_TEMPLATE, MLSP_TUNNEL_IPv6 C-Type = TBD

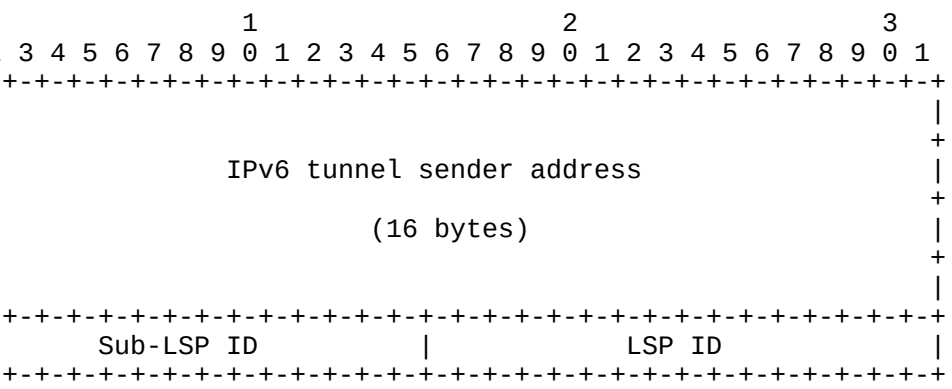


Figure 4: MLSP_TUNNEL_IPv6 Sender Template

3.1.2. MLSP_TUNNEL Filter Specification

The MLSP_TUNNEL Filter Specification also comes in two flavors, IPv4 and IPv6. The formats are identical to the IPv4 and IPv6 formats (respectively) of the MLSP_TUNNEL Sender Template. The Class numbers for both are FILTER SPECIFICATION, and the C-Types are (respectively) MLSP_TUNNEL_IPv4 (TBD) and MLSP_TUNNEL_IPv6 (TBD).

3.2. Label Allocation

A LSR S that receives Path messages for several sub-LSPs of the same MLSP from the same upstream LSR SHOULD allocate the same label for all the sub-LSPs. This simplifies load balancing for the aggregate traffic on those sub-LSPs. If the sub-LSPs are EB sub-LSPs, then S SHOULD allocate the same label for all EB sub-LSPs of the same MLSP that pass through S, regardless of which upstream LSR they come from. This allows S to load balance the aggregate traffic received on the MLSP, as all the MLSP traffic arrives at S with the same label. However, an LSR that can achieve the load balancing requirements independent of label allocation strategies is free to do so.

3.3. Bandwidth Accounting

Since MLSPs are traffic engineered, there needs to be strict bandwidth accounting, or admission control, on every link that an MLSP traverses. For non-EB sub-LSPs, this is straightforward, and analogous to regular TE LSPs. However, for EB sub-LSPs, two new procedures are needed, one for signaling bandwidth, and the other for admission control. First, for a given MLSP Z, an LSR X MUST ensure (via signaling) that the total incoming bandwidth of EB sub-LSPs of

MLSP Z is divided equally among all the downstream links of X which at least one of the EB sub-LSPs traverses. Second, LSR X MUST ensure that, for each upstream link of X, there is sufficient bandwidth to accommodate all EB sub-LSPs of MLSP Z that traverse that link.

Let's take the example of Figure 2, with MLSP Z having five EB sub-LSPs E1 to E5, and say that MLSP Z is configured with a bandwidth of 30Gbps. Here are some of the steps involved.

1. LSR A, being the ingress, has no upstream links. A has two downstream links, AL and AM. Three EB sub-LSPs of MLSP Z traverse AL, and two traverse AM. A MUST signal a total of 15Gbps for the sub-LSPs to L, and a total of 15Gbps for the sub-LSPs to M. The required bandwidth may be divided up among the sub-LSPs to L (similarly, to M) in any manner so long as the total is 15Gbps. For example, A can signal sub-LSP E1 with 15Gbps, and sub-LSPs E3 and E5 with 0 bandwidth.

2. LSR L has one upstream link AL with three EB sub-LSPs with a total bandwidth of 15Gbps. L MUST ensure that 15Gbps is available for the AL link. If this bandwidth is not available, L MUST send a PathErr on ALL of the EB sub-LSPs on the AL link. Let's assume that the AL link has sufficient bandwidth.

3. Next, it is up to L to decide how to divide the incoming 15Gbps among the three downstream EB sub-LSPs to S. Say L signals sub-LSP E1 with 15Gbps, and the others with 0 bandwidth.

4. LSR S has two upstream links: LS with three EB sub-LSPs with a total bandwidth of 15Gbps, and MS with two EB sub-LSPs with a total bandwidth of 15Gbps. S MUST ensure that 15Gbps is available for each of the LS and MS links. S has thus a total incoming bandwidth of 30Gbps on MLSP Z. S has to divide this equally among its downstream links SP, SQ and SR, yielding 10Gbps each. S MUST ensure that the total bandwidth requested on the SP link for sub-LSPs E1 and E4 is 10Gbps. S may choose to signal these sub-LSPs with 5Gbps each. Similarly for the SQ and SR links.

There are two important points to note here. One is that the bandwidth reservation (TSpec) for a given EB sub-LSP can (and usually will) change hop-by-hop. The second is that as new EB sub-LSPs are signaled for an MLSP, the bandwidth reservations for existing EB sub-LSPs belonging to the same MLSP may have to be updated. To minimize these updates, it is RECOMMENDED that the first EB sub-LSP on a link be signaled with the total required bandwidth (as far as is known), and later sub-LSPs on the same link be signaled with 0 bandwidth.

3.4. MLSP Data Plane Actions

Traffic intended to be sent over an MLSP is determined at the ingress LSR by means outside the scope of this document, and at transit LSRs by the label(s) assigned by the transit LSR to its upstream LSRs. In the case of non-EB sub-LSPs, this traffic is load balanced across downstream links in the ratio of the bandwidths of the sub-LSPs that comprise the MLSP. In the case of EB sub-LSPs, the traffic belonging to an MLSP from an upstream LSR (or better still, the aggregate traffic for the MLSP from all upstream LSRs) is load balanced equally among all downstream links.

As noted above, the overriding concern is that flows are mapped to the same downstream link (except when the MLSP or some constituent sub-LSPs are changing); this is typically done by hashing fields that define a flow, and mapping hash results to different downstream LSRs. Hash-based load balancing typically assumes that the numbers of flows is sufficiently large and the bandwidth per flow is reasonably well-balanced so that the results of hashing yields reasonable traffic distribution.

Entropy labels ([I-D.kompella-mpls-entropy-label] and [I-D.ietf-pwe3-fat-pw]) can be used to improve load balancing at intermediate nodes.

4. Security Considerations

This document introduces no new security concerns in the setup and signaling of LSPs using RSVP-TE, or in the use of the RSVP protocol. [RFC2205] specifies the message integrity mechanisms for RSVP signaling. These mechanisms apply to RSVP-TE signaling of MLSPs described in this document, and are highly recommended pending newer mechanisms for RSVP.

5. Acknowledgments

The author would like to thank the Routing Protocol group at Juniper Networks for their questions, comments and encouragement for this proposal. While many participated, special thanks go to Yakov Rekhter, John Drake and Rahul Aggarwal.

6. IANA Considerations

IANA is requested to assign two new C-Types for the Class "Sender Template", one for the "MLSP_TUNNEL_IPv4" Sender Template and one for the "MLSP_TUNNEL_IPv6" Sender Template.

IANA is also requested to assign two new C-Types for the Class "Filter Specification", one for the "MLSP_TUNNEL_IPv4" Filter Specification and one for the "MLSP_TUNNEL_IPv6" Filter Specification.

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.

[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.

[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.

7.2. Informative References

[RFC4655] Farrel, A., Vasseur, J., and J. Ash, "A Path Computation Element (PCE)-Based Architecture", RFC 4655, August 2006.

[RFC4972] Vasseur, JP., Leroux, JL., Yasukawa, S., Previdi, S., Psenak, P., and P. Mabbey, "Routing Extensions for Discovery of Multiprotocol (MPLS) Label Switch Router (LSR) Traffic Engineering (TE) Mesh Membership", RFC 4972, July 2007.

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

[RFC5152] Vasseur, JP., Ayyangar, A., and R. Zhang, "A Per-Domain Path Computation Method for Establishing Inter-Domain Traffic Engineering (TE) Label Switched Paths (LSPs)", RFC 5152, February 2008.

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

[RFC5286] Atlas, A. and A. Zinin, "Basic Specification for IP Fast Reroute: Loop-Free Alternates", RFC 5286, September 2008.

[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-07 (work in progress), July 2011.

[I-D.kompella-mpls-entropy-label]
Kompella, K., Drake, J., Amante, S., Henderickx, W., and L. Yong, "The Use of Entropy Labels in MPLS Forwarding", draft-kompella-mpls-entropy-label-02 (work in progress), March 2011.

Author's Address

Kireeti Kompella
Juniper Networks
1194 N. Mathilda Ave.
Sunnyvale, CA 94089
US

Email: kireeti@juniper.net

MPLS Working Group
Internet-Draft
Intended status: Informational
Expires: April 26, 2012

G. Liu
ZTE Corporation
Y. Weingarten
Nokia Siemens Networks
October 24, 2011

MPLS-TP protection for interconnected rings
draft-liu-mpls-tp-interconnected-ring-protection-00

Abstract

According to the ring protection Requirements in RFC 5654, Requirement 93 : When a network is constructed from interconnected rings, MPLS-TP MUST support recovery mechanisms that protect user data that traverses more than one ring. This includes the possibility of failure of the ring-interconnect nodes and links,so this document will describle all kinds of interconnected rings Scenario and a few possible solutions for recovery the failure of the ring-interconnect nodes and Links. .

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

Status of this Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79. This document may not be modified, and derivative works of it may not be created, and it may not be published except as an Internet-Draft.

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

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

This Internet-Draft will expire on April 26, 2012.

Table of Contents

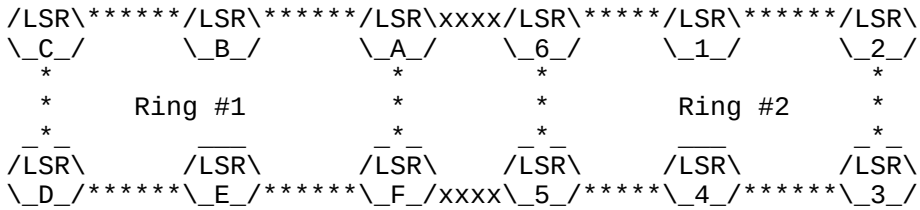
1. Introduction

This first version of the document will simply describle all kinds of interconnected rings scenario and a few protection solutions for the failure of the ring-interconnect nodes and links. For interconnected rings between two rings, there are mainly include three common interconnection scenario:

Dual-node interconnection - when the interconnected rings are interconnected by two nodes from each ring (see Figure 1);

Single-node interconnection - when the connection between the interconnected rings are through a single node (see Figure 2);

Chain of rings - when a series of rings are connected through interconnection nodes that are part of both interconnected rings (see Figure 3)



*** physical link
xxx interconnection link

Figure 1

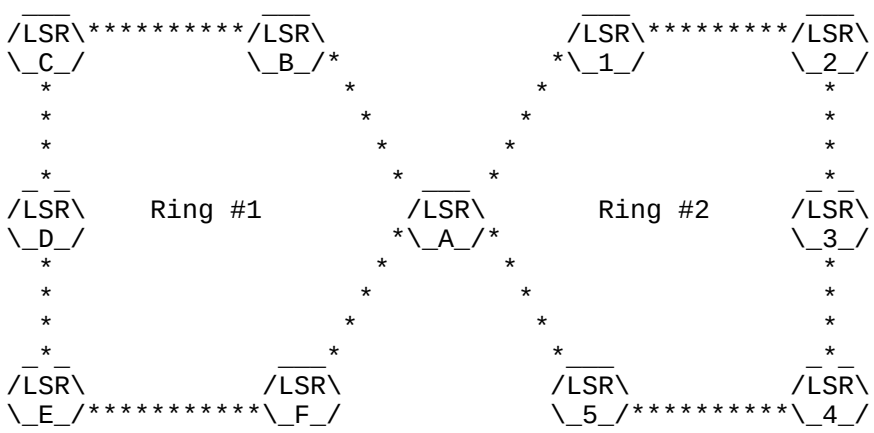

Ring #1
Ring #2
*** physical link

Figure 2

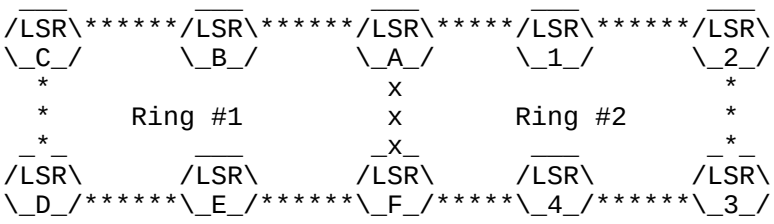

Ring #1
Ring #2
*** physical link
xxx shared link

Figure 3

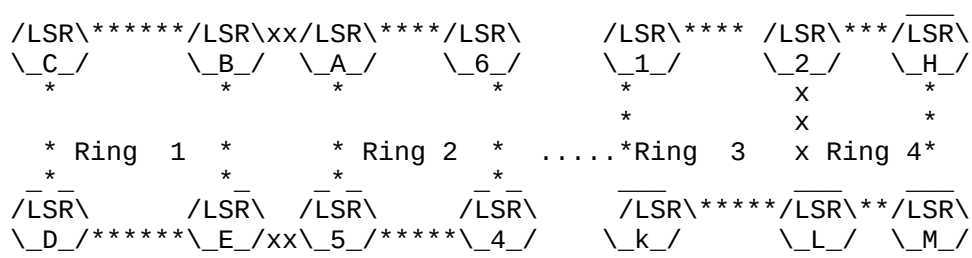
when a traffic traveres more than two rings.there are mainly the following mix interconnection scenarios:

Dual-node and single-node mix interconnection-when there not only exist two interconnected rings are interconnected by two nodes from each ring. but also there exist two interconnected rings are interconnected by single node(see figure 5);

Dual-node and chained mix interconnection-when there exist two interconnnected rings are interconnected by two nodes from each ring. in addtion, there still exist two interconnnected rings are interconnected by a common chained link(see figure 4);

single-node and chained mix interconnection-when there exist two interconnected rings are interconnected by single node, in addtion, there still exist two interconnected rings are interconnected by a common chained link(see figure 6);

Dual-node, single-node and chained mix interconnection-when there exist all three interconnection scenrios in the network domain including Dual-node interconnnection, single-node interconnection and chained interconnnection(see figure 7);



*** physical link
xxx interconnection link

Figure 4

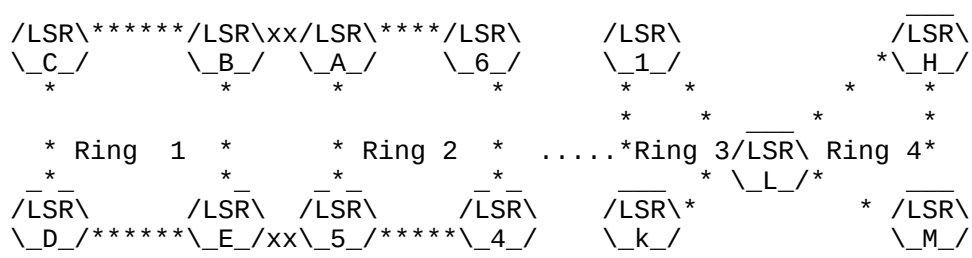


Figure 5

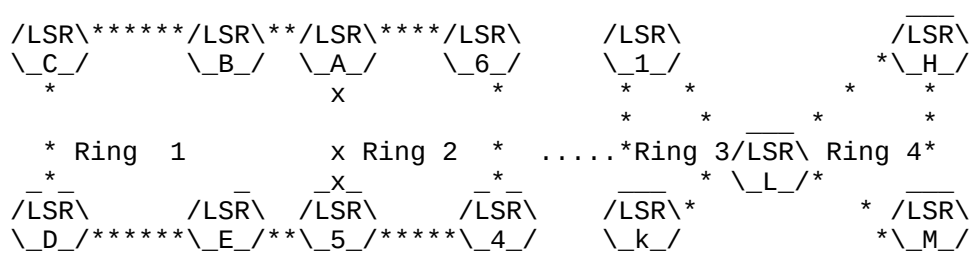


Figure 6

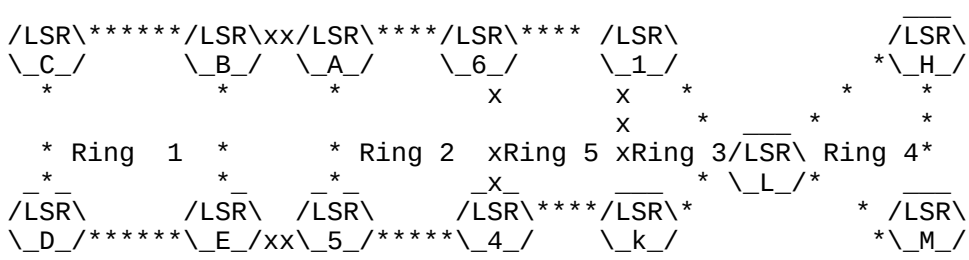


Figure 7

For a multi-ring service, it will be accross more than one ring just like above seven scenrios. if a failur happens on a multi-ring path, quickly recovery is necessary requirement for MPLS-TP network, so there are describles for recoverying the failure in the multi-ring interconnection sencrios in the following sections .

2. Conventions used in this document

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

OAM: Operations, Administration, Maintenance

LSP: Label Switched Path.

TLV: Type Length Value

P2MP:Point to Multi-Point

P2P:Point to Point

PSC:Protection Switching Coordination

SD:Signal Degrade

SF:Signal Fail

RDI:Remote Defect Indication

SPME:Sub-Path Maintenance Entity

MPLS-TP:Multi-Protocol Label Switching Transport Profile

ME: Maintenance Entity

MEP:MEG End Point

ACH: Associated Channel Header

CC-V: Contunuity Check-Verification;

3. recovery mechanisms

This section will describle recovery mechanisms that protect multi-ring traffics,which traver more than one ring in case of failure for all kinds of interconnection-ring scenrios;

3.1. recovery mechanism for Dual-node interconnected-ring

Under the interconnected-ring scenrios just as figure 1,multi-ring traffics will be transported by interconnection link(LSR C-LSR 6). when a failure happened on the interconnection link, if a segment protection path has been set up for the interconnection link , maybe apply 1:1 linear protection to protect the interconnection link faiure for interconnected-ring; or else, it maybe need end to end multi-ring path switch to protect the interconnection link failure. .

3.2. recovery mechanism for Single-node interconnected-ring

for the single-node interconnected-ring scenrio, As the interconnection node (LSR-A in Figure 2) is a single-point of failure , such an interconnection scheme should be avoided. .

3.3. recovery mechanism for Chained interconnected-ring

For the chained interconnected-ring scenrio, if the interconnection nodes(LSR-A and LSR-F) or the shared link(LSR-A-LSR-F) have failures, single ring protection solution can’t recovery the failure, so the affected multi-ring traffics maybe be protected by end to end protection path; .

3.4. recovery mechanism for Dual-node and Single-node mix interconnected-ring

for the mix interconnected-ring scenrios, each interconnection nodes or shared interconnection link will be protected by setting up segment protection path seperately. in addition, it may still use end to end multi-ring protection path to protect multiple interconnection nodes or shared interconnection link failure. .

3.5. recovery mechanism for Dual-node and Chained mix interconnected-ring

. for the mix interconnected-ring scenrios, each interconnection nodes or shared interconnection link will be protected by setting up segment protection path seperately. in addition, it may still use end to end multi-ring protection path to protect multiple interconnection nodes or shared interconnection link failure.

3.6. recovery mechanism for Single-node and Chained mix interconnected-ring

for the mix interconnected-ring scenrios, each interconnection nodes or shared interconnection link will be protected by setting up segment protection path seperately. in addition, it may still use end to end multi-ring protection path to protect multiple interconnection nodes or shared interconnection link failure.

3.7. recovery mechanism for Dual-node ,Single-node and Chained mix interconnected-ring

for the mix interconnected-ring scenrios, each interconnection nodes or shared interconnection link will be protected by setting up segment protection path seperately. in addition, it may still use end to end multi-ring protection path to protect multiple interconnection nodes or shared interconnection link failure.

4. Security Considerations

TBD

5. IANA Considerations

TBD.

6. Acknowledgments

TBD .

7. References

7.1. Normative References

[RFC 5654]
IETF, "IETF RFC5654(MPLS-TP requirement)", September 2009.

[RFC 5921]
IETF, "IETF RFC5654(MPLS-TP framework)", July 2010.

[RFC 6372]
N. Sprecher, A. Farrel, "Multiprotocol Label Switching Transport Profile Survivability Framework", September 2011.

7.2. Informative References

[MPLS-TP Linear protection]
S. Bryant, N. Sprecher, H. van Helvoort,A. Fulignoli Y. Weingarten, "MPLS transport profile Linear Protection", July 2010.

[MPLS-TP Ring Protection]
Y. Weingarten, "Multiprotocol Label Switching Transport Profile Ring Protection", Sep 2011.

7.3. URL References

[MPLS-TP-22]
IETF - ITU-T Joint Working Team, "", 2008, <http://www.example.com/dominator.html>.

Authors' Addresses

Liu guoman
ZTE Corporation
No.50, Ruanjian Road, Yuhuatai District
Nanjing 210012
P.R.China

Phone: +86 025 52871606
Email: liu.guoman@zte.com.cn

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

Phone: +972-9-775 1827
Email: yaacov.weingarten@nsn.com

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: April 27, 2012

J. Jeganathan
M. Konstantynowicz
H. Gredler
Juniper Networks
October 25, 2011

2547 egress PE Fast Failure Protection
draft-minto-2547-egress-node-fast-protection-00

Abstract

This document specifies a mechanism for protecting RFC2547 based VPN service against egress node failure. The mechanism enables local repair to be performed immediately upon a egress node failure. In particular, the router at point of local repair (PLR) can redirect VPN traffic to a protector to repair in the order of tens of milliseconds, achieving fast protection that is comparable to MPLS fast reroute.

Status of this Memo

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

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

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

This Internet-Draft will expire on April 27, 2012.

Copyright Notice

Table of Contents

1. Introduction

This document specifies a mechanism for protecting RFC2547 based VPN against egress PE failure. The procedures in this document are relevant only when a VPN site is multi-homed to two or more PEs. This is designed on the basis of MPLS context specific label switching [RFC 5331]. Fast-protection refers to the ability to provide local repair upon a failure in the order of tens of milliseconds, which is comparable to MPLS fast-reroute [RFC 4090]. This is achieved by establishing local protection as close to a failure as possible. Compared with the existing global repair mechanisms that rely on control plane convergence, these procedures can provide faster restoration for VPN traffic. However, they are intended to complement the global repair mechanisms, rather than replacing them in any way.

2. Specification of Requirements

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

3. Terminology

Protected PE: A PE which request protection for minimum one VPN prefix.

Protected prefix: A VPN prefix that required protection in case of Protected PE goes down.

Protector: A router which protect one or more VPN prefix when a Protected PE goes down.

BGP nexthop: A nexthop advertised in the BGP-Update for the VPN prefix by a BGP speaker.

VPN label: A label advertised by a BGP speaker for set of VPN prefixes. This label can be per-VRF label or per nexthop label or per prefix label.

Transport LSP: A LSP setup to BGP nexthop either by LDP or RSVP.

Alternative egress PE: A PE originates same IP prefix as Protected prefix in a same VPN.

VPN transport LSP: A Transport LSP that carries VPN traffic.

Context table: A context-specific label space routing table. This table is per is populated with VPN labels advertised by the protected-PE.

Context node: A stub router advertised into IGP by protected PE for a context-identifier.

4. Reference topology

This document refers to the following topologies to describe various roles and solution.

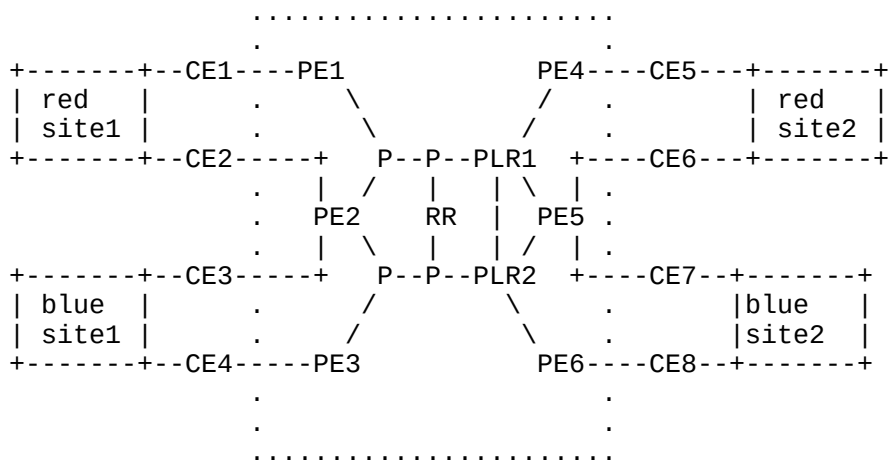


Figure 1

In Topology-1 two VPNs red and blue with two sites multihomed with PEs. Let assume blue and red VPN site2 prefixes required egress protection in case of PE5 goes down. PE5 is protected PE for site2 prefixes for both VPN. PE4 is alternate PE for red site2 prefixes. PE6 is alternate PE for blue site2 prefixes. For PE4 could act as protector for red VPN site2 and PE6 could acts as protector for blue VPN site2. This model is co-located protector model. RR could act as protector for both red and blue VPN site2. This is Centralized protector model (A PE protecting set of VPNs and not connected to any VPN site).

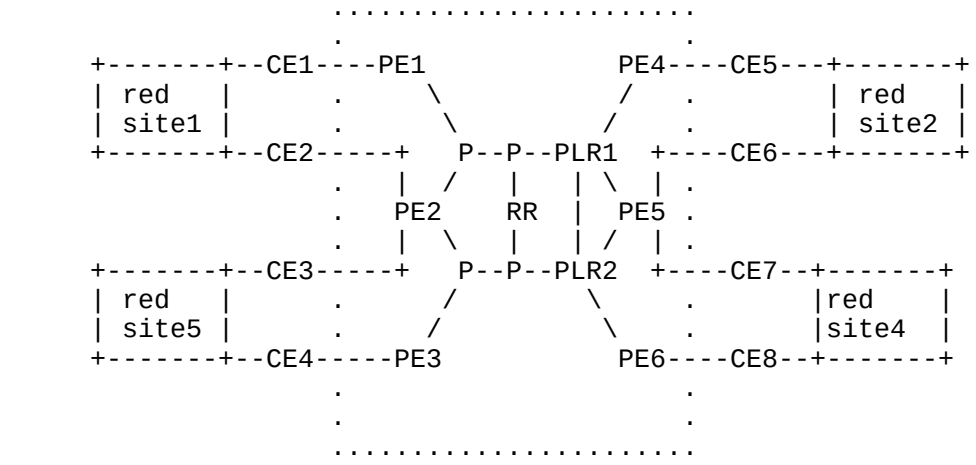


Figure 2

In Topology-2 has a VPNs red with four sites and multihomed with PEs. Let assume red VPN site2 and site4 prefixes required egress protection in case of PE5 goes down. PE5 is protected PE for site2, site4 prefixes for red VPN. PE4 is alternate PE for site2 prefixes. PE6 is alternate PE for site4 prefixes. Either PE4 or PE6 could act as protector. This is a slight variation of the co-located model.

5. Theory of Operation

The Egress PEs attached to multi-homed site export VPN prefixes with different route distinguisher, different nexthop but with same route target. The other PEs attached to other sites with same VPN import these route into VRF creates more than one path to multi-homed sites. When one egress PE goes down all VPN traffic towards the multihomed site moved to alternate egress PEs attached to the multi-homed site. This is done by ingress PE. The VPN traffic going via failed PE get dropped in penultimate hop router until ingress PE reroute VPN traffic. Even though connectivity of multi-homed site is not bound to an egress PE the transport LSP bind to egress PE. As result of down transport LSP VPN traffic getting dropped in P router. This document specifies a mechanism that repair VPN traffic at point of failure (typically a P router which penultimate hop of the transport LSP) and still keep P router unaware of the VPN information with the help of protector (a new role). The PLR (point of local repair) send VPN traffic to protector through bypass LSP incase of egress PE failure. This protector send VPN traffic received from PLR to the alternative egress PE until the ingress reroute traffic to alternate egress PE.

5.1. Protector and Protection Models

Protector, is a new role for the egress PE failure local repair. This protector role could be played by a PE(alternate egress PE) or any other nodes which participates in VPN control plane for VPN prefixes that required egress node protection. Hence, there are two protection models based on the location and role of a protector.

5.1.1. Co-located protector

In this model, the protector is a alternate egress PE for a protected prefix. It is co-located with the alternate PE for the protected prefix, and it has a direct connection to the multi-homed site that originate the protected prefix. In the event of an egress node failure, the protector receives traffic from the PLR, and sends the traffic to the multi-homed site. In the topology-1 PE4 could act as protector for red VPN site2 and PE6 could acts as protector for blue VPN site2. This model is co-located protector model. RR could act as protector for both red and blue VPN site2. This is Centralized protector model (A PE protecting set of VPNs and not connected to any VPN site).

A slight variant of this model is, protector is not alternate PE for a protected prefix but has same VRF. In the topology-2 either PE4 or PE6 could act as protector. This is example for the above model.

5.1.2. Centralized protector

In this model, the protector serves as a centralized protector MAY NOT have a direct connection to multi-homed site. This model can be played by existing PEs or other PEs. In the event of an egress PE failure, protector MUST send the traffic to a alternate egress PE with VPN label advertised alternate egress PE for the prefix which in turn sends the traffic to the multi-homed site. In the topology-1 RR could act as protector for both red and blue VPN site2. This is Centralized protector model (A PE protecting set of VPNs and not connected to any VPN site).

A network MAY use either protection model or a combination of both, depending on requirements.

5.2. Context Identifier and VPN prefixes.

The context-identifier is an IP address that is either globally unique or unique in the private address space of the routing domain. In Egress PE each VPN prefix is assigned to context-identifier. The granularity of a context identifier is {Egress PE, VPN prefix} tuple. However, a given context identifier MAY be assigned to one or

multiple VPN prefix.

Possible context identifier assignments are

- Unique context-identifier for all VPN prefixes, both VPN-IPv4 and VPN-IPv6.
- Unique context-identifier per address family.
- Unique context-identifier per site for all VPN prefixes, both VPN-IPv4 and VPN-IPv6.
- Unique context-identifier per site per address family.
- Unique context-identifier per CE address (nexthop).
- Unique context identifier for each prefix.

The first one is coarsest granularity of a context identifier and the last one is finest granularity of a context identifier. While all of the above options are possible in principle, their practical usage is likely to vary widely, as not all of them may be of practical usage. A given context identifier MUST NOT be used by more than one protected PE. The egress PE that required protection for a VPN prefix MUST put context-identifier as nexthop in BGP update. This context-identifier as nexthop indicates to protector that this prefix need protection. For e.g. In topology 1 PE5(protected PE) advertise VPN prefixes with context-identifier as BGP nexthop.

5.3. Context Identifier Advertisement by IGP

IGP MUST advertise context identifiers to allow computation of TE paths for bypass LSPs and VPN transport LSPs that are destined for context identifiers. Context identifiers MUST be advertised a stub router in IGP and TE. Advertised as a stub router allow operator to deploy egress protection without upgrading all P routers.

A protected PE MUST advertise a context identifier as a stub router to TE domain and in IGP. Also Protected PE MUST advertise a link to the stub router.

A protector MUST advertise link to stub router advertised by protected PE in IGP and TE.

5.3.1. Context-identifier advertised as stub router.

Context-identifier advertised as stub router required two parts. A node representation (context-node) and links to the node. The

protected PE and protector advertise link to context-node and protected PE advertise context-node.

The protected PE will advertise context-node in to IGP. The router-id of the context-node is context-identifier. The system-ID is derived from the context-identifier with BCD encoding. The resulting system-ID MUST be unique with in IGP routing domain. Context-node advertised with two unnumbered transit links with MAX routable link metric to protected PE and protector. For TE these unnumbered links advertised with zero bandwidth and MAX TE metric. Other TE characteristic of TE links could be configured to advertise. The router-ID or system-ID of the protector could be dynamically learned from the IGP link state database or could be configured in protected PE.

Protected PE MUST advertise unnumbered transit link with metric 1 and TE metric 1 to context-node. Protector MUST advertise unnumbered transit link with maximum routable link metric and maximum TE metric to the context-node. Other TE characteristic of the links could be configured and advertised in to TE.

5.3.1.1. ISIS context-node

Only zeroth fragment of the context-node is only valid. All Other fragments SHOULD be ignored. Zeroth fragment MUST include area address TLV and MAY include hostname TLV.

The set of area addresses advertised MUST be a subset of the set of Area Addresses advertised in the protected LSP number zero at the corresponding level. Preferably, the advertisement SHOULD be syntactically identical to that included in the normal LSP number zero at the corresponding level. The hostname could be set as <context-identifier+ protected PE hostname>.

The Overload (OL) MUST be set to 1. The Attached (ATT), and Partition Repair (P) bits MUST be set to 0.

5.3.1.2. OSPF context-node

The advertising router and Link State ID of router LSA MUST be context-identifier. All options bits in router LSA MUST be set to zero. The number of links MUST be 2.

5.4. Forwarding State on Protector PE

A protector maintain the forwarding state in context-specific label spaces on a per protected PE basis. In particular, the protector MUST learn the VPN label by participating the VPN routing and also

MUST keep all routes associated with VPN it required to protect.

For each VPN label with an associated context-identifier protector MUST map the context identifier to a context-specific label space [RFC 5331], and program the VPN label in that label space in forwarding plane. For each VPN prefix that required protection programmed in the forwarding plane with BGP nexthop to alternate egress PE. This VPN label in the context-specific label space identify the layer-3 forwarding table that need to lookup to send it alternate egress PE. The protector MAY maintain only VPN prefix originated with-in the multi-homed site for given {egress PE, VPN}. These VPN labels in context table and VPN context table will not be used in forwarding after ingress reroute the traffic to alternative PE. Protector MUST delete VPN label and the VPN context table after ingress reroute the traffic. This shall be achieved with a timer. This timer default value is 180 seconds.

5.4.1. Alternate egress PE for protected prefix.

Any route with BGP nexthop which has the following properties

Exact matching route-target set (RD may be different)

Exact matching Prefix part (not RD)

will be eligible as alternate egress PE for prefix.

5.5. Bypass LSP

An LSP MUST be either manually or automatically provisioned on a PLR to enable the PLR to send traffic to a protector, in the event of an egress PE failure. This LSP is referred to as a bypass LSP. The bypass LSP MUST be a LSP with ultimate hop popping (UHP) [RFC 3031]. That is, the protector MUST assign an un-reserved label to the bypass LSP. When the protector PE receives VPN packets on the bypass LSP from a PLR, it MUST rely on the bypass LSP’s UHP label to determine the context-specific label space to forward the packets.

5.5.1. RSVP-TE Signaled Bypass LSP and Backup LSP

If a bypass LSP is an RSVP-TE signaled LSP, its destination MUST be the context identifier of the protected VPN prefix. The path taken by the bypass LSP MAY be statically configured or dynamically computed by CSPF. The signaling of the bypass LSP MUST be triggered by the "local protection desired" and "node protection desired" bits in SESSION_ATTRIBUTE of Path message of the transport LSP [RFC 2205, RFC 3209, RFC 4090]. After the bypass LSP is established, the PLR MUST set the "local protection available" and "node protection" bits

in RRO of Resv message of the transport LSP. The protector MUST terminate the backup LSP as an egress. Once the local repair takes effect, the PLR MUST set the "local protection in use" bit in RRO of Resv message of the transport LSP.

5.5.2. LDP Signaled Bypass LSP

If it is LDP LSP then LDP FEC for this LSP MUST be the context identifier of the protected segment. Prefix LFA with node protection can be used for bypass LSP computation.

6. Egress node Failure

This section summarizes the procedures egress protection described above section for completeness. A Egress PE and a protector both advertise the context identifier of a protected prefixes in IGP as a stub link or stub router, with the egress PE advertising a lower metric and protector with maximum metric. The PLR establishes a UHP bypass LSP to the protector. The destination address of the bypass LSP is the context identifier. The protector programs forwarding state in such a way that packets received on the bypass LSP will be forwarded based on VPN label in the context table, and prefix lookup in VPN context table. The context table identified by the UHP label of the bypass LSP, i.e. the context identifier.

When the penultimate Hop router receives a VPN packet from the MPLS network, if the egress PE is down, the PLR tunnels the packet through the bypass LSP to the protector. The protector PE identifies the forwarding context of the egress PE based on the top label of the packet which is the UHP label of the bypass LSP. Then forwards protector the packet based on a second label lookup in the protected PE’s context label space followed by layer-3 lookup in the VPN context table. These UHP label, context table label and layer-3 lookup results in forwarding the packet to the site or send it to alternate egress PE based on protector model.

For E.g. In topology-1 RR is act as Protector and PE5 required protection for red, blue site2 prefixes. As red, blue site2 VPN prefixes advertised with context-identifier, the protector set up the forwarding table for prefixes from site2 with alternative egress PE as nexthop. When PLR detects PE5 failure it send to protector through bypass LSP. In protector the top label identify the context space table. VPN label in the context table identify the VPN layer-3 forwarding table with contains site2 prefixes with alternate PE as nexthop. A Layer-3 lookup gives mpls path to alternate egress PE and protector forward packet to alternate egress PE and reach to the site2.

7. Security Considerations

The security considerations discussed in RFC 5036, RFC 5331, RFC 3209, and RFC 4090 apply to this document.

8. Acknowledgements

This document leverages work done by Yakov Rekhter and several others on LSP tail-end protection. Thanks to Nischal Sheth, Nitin Bahadur, Yimin shen for their contribution.

9. References

9.1. Normative References

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

[RFC4364] Rekhter, Y. and E. Rosen, "BGP/MPLS IP Virtual Private Networks (VPNs)", RFC 4364, February 2006.

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

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

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

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

[RFC3471] Berger, L., "Generalized Multi-Protocol Label Switching (GMPLS) Signaling Functional Description", RFC 3471, January 2003.

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

[LDP-UPSTREAM] Aggarwal, R. and J. Roux, "MPLS Upstream Label Assignment

for LDP", draft-ietf-mpls-ldp-upstream (work in progress), 2011.

[RSVP-NON-PHP-OOB]
Ali, A., Swallow, Z., and R. Aggarwal, "Non PHP Behavior and out-of-band mapping for RSVP-TE LSPs", draft-ietf-mpls-rsvp-te-no-php-oob-mapping (work in progress), 2011.

9.2. Informative References

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

[RFC5286] Atlas, A., Ed. and A. Zinin, Ed., "Basic Specification for IP Fast Reroute: Loop-Free Alternates", RFC 5920, September 2008.

[RFC5714] Shand, M. and S. Bryant, "IP Fast Reroute Framework", RFC 5714, January 2010.

Authors' Addresses

Jeyananth Minto Jeganathan
Juniper Networks
1194 N Mathilda Avenue
Sunnyvale, CA 94089
USA

Email: minto@juniper.net

Maciek Konstantynowicz
Juniper Networks
1194 N Mathilda Avenue
Sunnyvale, CA 94089
USA

Email: maciek@juniper.net

Hannes Gredler
Juniper Networks
1194 N Mathilda Avenue
Sunnyvale, CA 94089
USA

Email: hannes@juniper.net

Juniper Networks

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

not be created outside the IETF Standards Process, except to format it for publication as an RFC or to translate it into languages other than English.

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

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

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

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

This Internet-Draft will expire on January 11, 2012.

Abstract

Shared mesh protection is a common protection and recovery mechanism in transport networks, where multiple paths can share the same set of network resources for protection purposes.

In the context of MPLS-TP, it has been explicitly requested as a part of the overall solution (Req. 67, 68 and 69 in RFC5654 [1]).

It's important to note that each MPLS-TP LSP may be associated with transport network resources. In event of network failure, it may require explicit activation on the protecting paths before switching user traffic over.

In this memo, we define a lightweight signaling mechanism for protecting path activation in shared mesh protection-enabled MPLS-TP networks.

Table of Contents

1. Introduction

Shared mesh protection is a common 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.

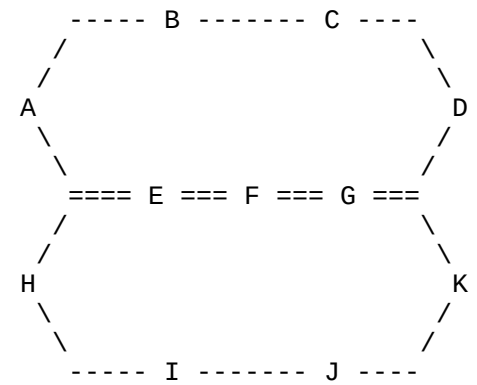
The activation signaling defined in this memo is to support warm standby in the context of MPLS-TP.

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:

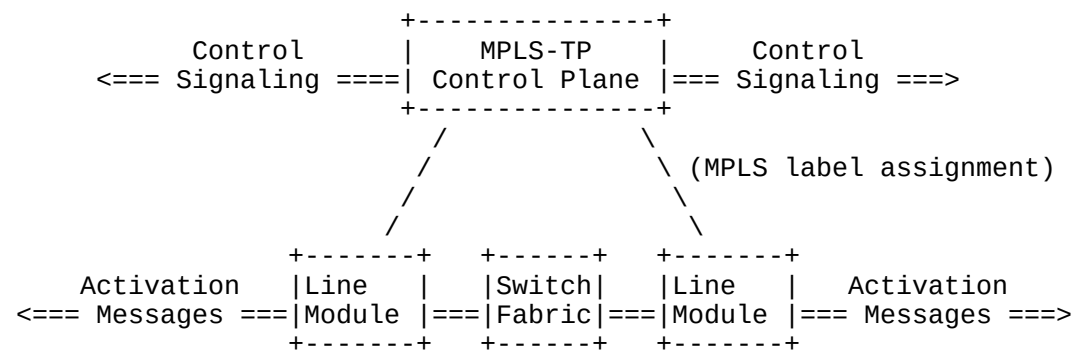
1. Preemption: Protecting paths X' and Y' may share the same resources on node E, F or G due to resource constraints. Y' has higher priority than that of X' In the previous example, X' is up and running. When there is a link outage on I-J, H can activate its protecting path Y' On E, F or G, Y' can take 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.

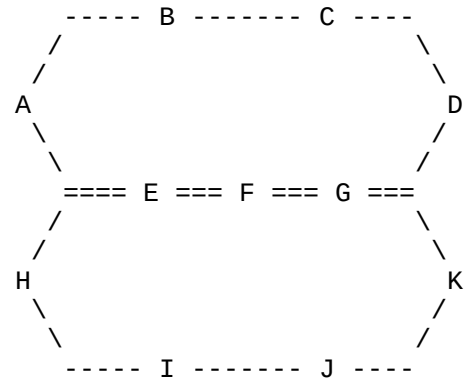
To achieve the tight end-to-end switch-over budget, it's possible to implement the entire activation procedure with hardware-assistance (such as in FPGA or ASIC).

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.

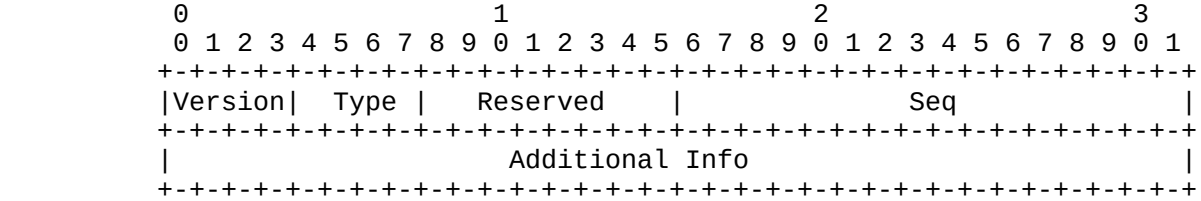
6. Protocol Definition

6.1. Activation Messages

The activation requires the following messages:

- ENABLE: this is initiated by the headend nodes to activate a protecting path
- DISABLE: this is initiated by the headend nodes to disable a protecting path and free the associated network resources
- GET: this is initiated by the headend to gather resource availability information on a particular protecting path
- NOTIFY: this is initiated by the intermediate nodes and terminate on the headend nodes to report preemption or protection failure conditions
- STATUS: this is the acknowledgement message for ENABLE, DISABLE, GET, and NOTIFY messages, and contains the relevant status information

Each activation message has the following format:



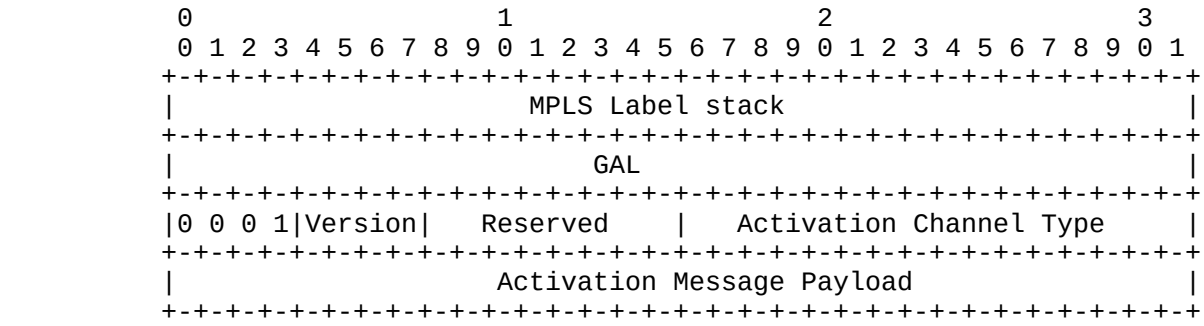
- Version: 1
- Type:

- 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

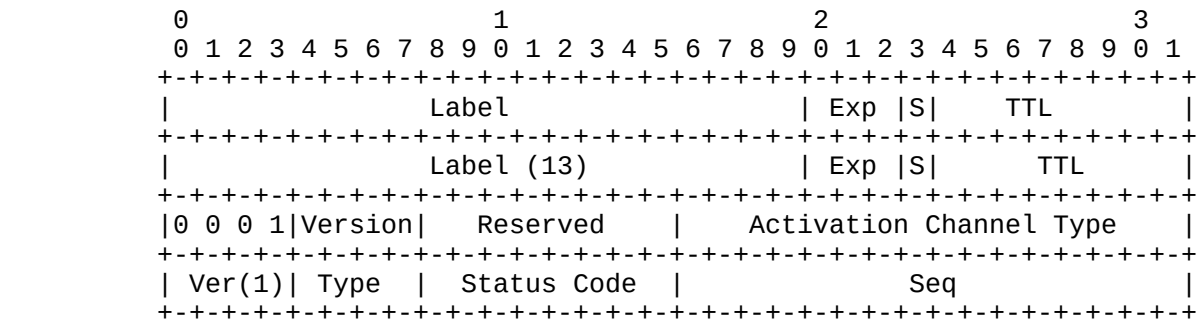
6.2. Message Encapsulation

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:

- On receive, if the message has label TTL = 0, the node must drop the packet without further processing
- The receiving node must always decrement the label TTL value by one. If TTL = 0 after the decrement, the node must process the message. Otherwise, the node must forward the message without further processing (unless, of course, the node is headend or tailend)
- On transmission, the node will adjust the TTL value. For hop-by-hop messages, TTL = 1. Otherwise, TTL = 0xFF, by default.

7. Processing Rules

7.1. Enable a protecting path

Upon the detection of network failure on a working path, the headend node identifies the corresponding MPLS-TP label and initiates the protection switching by sending an ENABLE message.

ENABLE messages always use MPLS label TTL = 1 to force hop-by-hop process. Upon reception, a next-hop node will locate the corresponding path and activate the path.

If the Enable message is received on an intermediate node, due to label TTL expiry, the message is processed and then propagated to the next hop of the MPLS TP LSP, by setting the MPLS TP label TTL = 1. The intermediate node may NOT respond back to the headend node with STATUS message.

The headend node will declare the success of the activation only when it gets a positive reply from the tailend node. This requires that the tailend nodes must reply STATUS messages to the headend nodes in all cases.

If the headend node is not receiving the acknowledgement within a time internal, it will retransmit another ENABLE message with a different Seq number.

If the headend node is not receiving a positive reply within a longer time interval, it will declare activation failure.

If an intermediate node cannot activate a protecting path, it will reply an NOTIFY message to report failure. When the headend node receives a NOTIFY message for failure, it must initiate DISABLE messages to clean up networks resources on all the relevant nodes on the path.

7.2. Disable a protecting path

The headend removes the network resources on a path by sending DISABLE messages.

In the message, the MPLS label represents the path to be de-activated. The MPLS TTL is one to force hop-by-hop processing.

Upon reception, a node will de-activate the path, by freeing the resources from the data-plane.

As a part of the clean-up procedure, each DISABLE message must traverse through and be processed on all the nodes of the corresponding path. When the DISABLE message reaches to the tailend node, the tailend is required to reply with a STATUS message to the headend.

The de-activation process is complete when the headend receives the corresponding STATUS message from the tailend.

7.3. Get protecting path status

The operators have the option to trigger GET messages from the headend to check on the protecting path periodically or on-demand. The process procedure on each node is very similar to that of ENABLE messages on the intermediate nodes, except the GET messages should not trigger any network resource re-programming.

Upon reception, the node will check the availability of resources.

If the resource is no longer available, the node will reply a NOTIFY with error conditions.

7.4. Acknowledgement with STATUS

The STATUS message is the acknowledgement packet to all messages, and may be generated by any node in the network.

Each STATUS message must use the same sequence number as the corresponding message (ENABLE, DISABLE, GET and NOTIFY).

When replying to headend, the tailend nodes must originate STATUS messages with a large MPLS TTL value (0xff, by default).

7.5. Preemption

The preemption operation typically takes place when processing an ENABLE message.

If the activating network resources have been used by another path and carrying user traffic, the node needs to compare the priority levels.

If the existing path has higher priority, the node needs to reject the ENABLE message by sending a STATUS message to the corresponding headend to inform the unavailability of network resources.

If the new path has higher priority, the node will reallocate the resource to the new path, and send an NOTIFY message to old path’s headend node to inform about the preemption.

8. Security Consideration

The protection activation takes place in a controlled networking environment. Nevertheless, it is expected that the edge nodes will encapsulate and transport external traffic into separated tunnels, and the intermediate nodes will never have to process them.

9. IANA Considerations

Activation messages are encapsulated in MPLS-TP with a specific GACH channel type that needs to be assigned by IANA.

10. Normative References

[1] RFC 5654: Requirements of an MPLS Transport Profile, B. Niven-Jenkins, D. Brungard, M. Betts, N. Sprecher, S. Ueno, September 2009

[2] IETF draft, Multiprotocol Label Switching Transport Profile Survivability Framework (draft-ietf-mpls-tp-survive-fwk-06.txt), N. Sprecher, A. Farrel, June 2010

[3] RFC5586 - Vigoureux,, M., Bocci, M., Swallow, G., Aggarwal, R., and D. Ward, "MPLS Generic Associated Channel", May 2009.

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

[5] Crocker, D. and Overell, P.(Editors), "Augmented BNF for Syntax Specifications: ABNF", RFC 2234, Internet Mail Consortium and Demon Internet Ltd., November 1997.

11. Acknowledgments

Authors like to thank Eric Osborne, Lou Berger, Nabil Bitar and Deborah Brungard for detailed feedback on the earlier work, and the guidance and recommendation for this proposal.

We also thank Maneesh Jain, Mohit Misra, Yalin Wang, Ted Sprague, Ann Gui and Tony Jorgenson for discussion on network operation, feasibility and implementation methodology.

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: Standard
Expires: January 2012

Pranjal Kumar Dutta
Alcatel-Lucent

Giles Heron
Cisco Systems

Thomas Nadeau
CA Technologies

July 3, 2011

Targeted LDP Hello Reduction
draft-pdutta-mpls-tldp-hello-reduce-02.txt

Status of this Memo

Abstract

Targeted LDP Hellos are used for establishing adjacencies with non-directly connected peers. After an LDP session is established to a targeted peer, the session Keepalives are sufficient to notify the intent of an LSR to maintain its adjacency with the peer. This document proposes a mechanism to turn off Targeted LDP Hellos after LDP session is established to a peer.

Table of Contents

1. Introduction

LDP Hello messages are exchanged as part of the LDP discovery mechanism [RFC5036]. There are two types of LDP discovery mechanism described in [RFC5036] - Basic Discovery and Extended Discovery.

To engage in LDP Basic Discovery on an interface, an LSR periodically sends LDP Link Hellos out the interface to the well-known LDP discovery port for the "all routers on this subnet" group multicast address. Receipt of an LDP Link Hello on an interface, identifies a hello adjacency with a potential LDP peer reachable at the link level on the interface. Thus an LSR may establish hello adjacency with multiple peers discovered over a single interface and must continue to transmit hellos at regular intervals even after hello adjacency is established to a peer.

Extended discovery is used to support LDP sessions between non-directly connected LSRs. An LDP Targeted Hello is sent to a specific address rather than to the "all routers" group multicast address for the ongoing interface. Receipt of a LDP Targeted Hello indentifies a hello adjacency with a potential LDP peer at network level.

In Extended discovery there can be only one Targeted Hello Adjacency between two peers. Note that throughout this document "peer" means the LDP LSR designated by a unique LDP Identifier. Once the LDP session is operational between two targeted LDP peers, periodic session Keepalives are used to maintain the LDP session. After the session is operational the periodic Targeted Hellos between the LSRs become redundant, as session Keepalives in turn serves the intent of each LSR to maintain its adjacency to its peer.

When an LSR maintains a large number of LDP sessions (in thousands) to targeted peers, it is an additional burden to send and receive Targeted Hellos for all peers at periodic intervals. In MPLS deployments at access or mobility backhaul, there can be very large volume of LDP sessions with targeted LDP adjacencies to each base station. Moreover additional mechanisms such as centralized BFD [BFD] may be used to track liveliness of ldp sessions.

Another problem with targeted hello adjacency arises is Denial Of Service (DoS)_attacks. It is possible that existing hello adjacencies can get lost due to DoS attack on LDP Hello receiver by spurious hello packets. Unlike TCP sessions it is not always possible to provide per peer protection for UDP based hellos. Implementations can use methods to protect existing adjacencies while throttling spurious adjacencies but such methods may not be available in low cost MPLS devices in access. So it is important to avoid dependency on targeted LDP hellos on session maintenance as far as possible.

This document proposes an optional mechanism to turn off Targeted LDP Hellos after a LDP session is established to a targeted peer, without changes in the procedures defined in [RFC5036].

2. Conventions used in this document

INFO (REMOVE): INCLUDE THIS SECTION OR PORTIONS THEREOF IF DESIRED

In examples, "C:" and "S:" indicate lines sent by the client and server respectively.

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

This document uses the terminology defined in [RFC3031] and [RFC5036].

4. Targeted LDP Hello Reduction Procedure

The Targeted LDP Hello Reduction procedure uses the existing Common Hello Parameters TLV defined in [RFC5036]. Figure 1. shows the encoding of the TLV from [RFC5036] for reference.

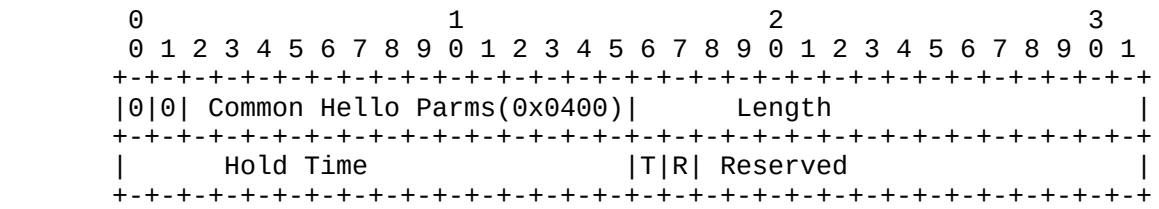


Figure 1. Common Hello Parameters TLV.

By definition in [RFC5036], a value of 0 means use the default, which is 45 seconds for Targeted Hellos. A value of 0xffff means infinite.

The procedure to be followed for Targeted LDP Hello Reduction between a pair of LSRs is as follows:

1. An LSR starts transmitting periodic targeted hellos to its peer in order to establish the targeted hello adjacency. Each LSR proposes its configured hello hold time in the Common Hello Parameters TLV in its hello message to the peer. The hold time used between a pair of LSRs is the minimum of the hold times proposed in their Hellos.

2. If the Hello is acceptable by receiving LSR it establishes targeted hello adjacency with the source LSR. Establishment of Hello adjacency establishes the LDP session between peering LSRs.

3. After the LDP session is ESTABLISHED [RFC5036], each LSR MAY advertise hello holdtime value of 0xffff in the Common Hello Parameters TLV. Thus after the session is ESTABLISHED, the hello hold time between the LSRs gets negotiated to infinite. An LSR MAY implement a locally configurable "tolerance" - the number of Targeted LDP Hellos to be advertised with infinite hold time after the LDP session is ESTABLISHED.

4. If the LDP session between two LSRs fails leading to tearing down of adjacency, then each LSR reverts to advertising their configured hello hold time and repeats procedure 1 to 3.

It is RECOMMENDED that each peering LSR implements the Targeted LDP Hello Reduction procedure; otherwise negotiated hello hold time between the LSRs does not fall back to the infinite hold time in step 3.

Note that it is not mandatory to advertise infinite hold time after session is established but can be any value that is significantly larger than configured hello hold time. It is RECOMMENDED to advertise Inifinite holdtime after session setup to derive maximum advantage from the procedure described above.

5. Security Considerations

- Control plane aspects
  - LDP security (authentication) methods as described in [RFC5036] is applicable here.

- Data plane aspects
  - This specification does not have any impact on the MPLS forwarding plane setup by LDP.

6. IANA Considerations

This document does not require any IANA consideration.

7. Conclusion

The method proposed in the document reduces significant burden on an LDP LSR that maintains Targeted LDP sessions to a large number (in thousands) of peers. Further, if BFD [BFD][BFD-MHOP] is used for tracking connectivity to peers it is desirable to turn off Targeted LDP hellos after the LDP session is setup.

8. References

8.1. Normative References

[RFC5036] Andersson, L., et al. "LDP Specification", RFC5036, October 2007.

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

8.2. Informative References

[RFC3031] Rosen, E., et al. "Multiprotocol Label Switching Architecture", RFC 3031, January 2001.

[BFD] Katz, D., et al. "Bidirectional Forwarding Detection", draft-ietf-bfd-base-011.txt, January 2010.

[BFD-MHOP] Katz, D., et al. "BFD for Multihop Paths", draft-ietf-bfd-multihop-09.txt, January 2010.

9. Acknowledgments

The authors would like acknowledge the comments and suggestions from Wim Henderickx, Vach Kompella, Florin Balus, Mustapha Aissaoui, Mathew Bocci and Paul Kwok.

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

Authors' Addresses

Pranjal Kumar Dutta
701 E Middlefield Road,
Mountain View, CA 94043.
USA.
Email: pranjal.dutta@alcatel-lucent.com

Giles Heron
Cisco Systems
Email: giheron@cisco.com

Thomas D. Nadeau
CA Technologies
273 Corporate Drive, Portsmouth, NH, USA

Email: thomas.nadeau@ca.com

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

Table of Contents

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-mpls-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-mpls-ldp-upstream-10.txt, Work in Progress, February 2011.

[RFC5561] Thomas, B., Raza, K., Aggarwal, S., Aggarwal, R., and Le Roux, JL., "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: April 28, 2012

Sam Aldrin
Huawei Technologies
M.Venkatesan
Kannan KV Sampath
Aricent Group
Thomas D. Nadeau
CA Technologies

October 26, 2011

BFD Management Information Base (MIB) extensions for MPLS and MPLS-TP Networks draft-vkst-bfd-mpls-mib-00

Abstract

This draft defines a portion of the Management Information Base (MIB) for use with network management protocols in the Internet community. In particular, it extends the BFD Management Information Base BFD-STD-MIB and describes the managed objects for modeling Bidirectional Forwarding Detection (BFD) protocol for MPLS and MPLS-TP networks.

Status of this Memo

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

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

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

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

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

This Internet-Draft will expire on April 28, 2012.

Table of Contents

1 Introduction

Current MIB for BFD as defined by BFD-STD-MIB is used for neighbor monitoring in IP networks. The BFD session association to the neighbors being monitored is done using the source and destination IP addresses of the neighbors configured using the respective MIB objects.

To monitor MPLS/MPLS-TP paths like tunnels or Pseudowires, there is a necessity to identify or associate the BFD session to those paths.

This memo defines an portion of the Management Information Base (MIB) for use with network management protocols in the Internet community. In particular, it extends the BFD Management Information Base BFD-STD-MIB and describes the managed objects to configure and/or monitor Bidirectional Forwarding Detection (BFD) protocol for MPLS [BFD-MPLS] and MPLS-TP networks [MPLS-TP-CC-CV-RDI].

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 adopts the definitions, acronyms and mechanisms described in [BFD], [BFD-1HOP], [BFD-MH], [BFD-MPLS], [MPLS-TP-CC-CV-RDI]. Unless otherwise stated, the mechanisms described therein will not be re-described here.

4. Acronyms

BFD: Bidirectional Forwarding Detection
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
ME: Maintenance Entity
MEG: Maintenance Entity Group
MEP: Maintenance Entity End-Point
PW: Pseudowire
TP: Transport Profile

5. Brief description of MIB Objects

The objects described in this section support the functionality described in documents [BFD-MPLS] and [MPLS-TP-CC-CV-RDI]. The objects are defined as an extension to the BFD base MIB defined by BFD-STD-MIB.

5.1. Extensions to the BFD session table (bfdSessionTable)

The BFD session table used to identify a BFD session between a pair of nodes, as defined in BFD-STD-MIB, is extended with managed objects to achieve the required functionality in MPLS and MPLS-TP networks as described below:

1. SessionRole - Active/Passive role specification for the BFD session configured on the node. Either end of a BFD session can be configured as Active/Passive to determine which end starts transmitting the BFD control packets.

2. SessionMode - Defines the mode in which BFD session is running, defined as below:
   i. CC - Only Continuity Check and RDI functionality is performed.

   ii. CV - Provides for Continuity Check, Connectivity Verification and RDI functionalities to be supported.

3. Timer Negotiation Flag - Provides for timer negotiation to be enabled or disabled. This object can be used to tune the detection of period-misconfiguration.

4. Map Type - Indicates the type of the path being monitored by

the BFD session.
This object can take the following values:

For BFD session over MPLS based paths:

nonTeIpv4 (1) - BFD session configured for Non-TE
Ipv4 path
nonTeIpv6 (2) - BFD session configured for Non-TE
Ipv6 path
teIpv4 (3) - BFD session configured for a TE
Ipv4 path
teIpv6 (4) - BFD session configured for a TE
Ipv6 path
pw (5) - BFD session configured for a PW

For MPLS-TP based paths:

mep (6) - BFD session configured for an MPLS-TP path
(Bidirectional tunnel, PW or Sections) will map to
the corresponding maintenance entity.

5. Map Pointer

A Row Pointer object which can be used to point to the first accessible object in the respective instance of the table entry identifying the path being monitored (mplsXCEntry/mplsTunnelEntry/ pwEntry respectively for LSP/Tunnel/PW).

For NON-TE LSP, the Map pointer points to the corresponding mplsXCEntry.

For TE based tunnel, the Map pointer points to the corresponding instance of the mplsTunnelEntry.

For PW, this object points to the corresponding instance of pwEntry.

For MPLS-TP paths, this object points to the corresponding instance of mplsOamIdMeEntry configured to monitor the MPLS-TP path associated with the BFD session.

6. Usage of existing object bfdSessType:

Additionally existing object "bfdSessType" in the base MIB can be used with the already defined value multiHopOutOfBandSignaling(3) to specify an OOB (Out of band) mechanism [E.g. LSP Ping] for bootstrapping the BFD session.

5.2. Example of BFD session configuration

This section provides an example BFD session configuration for an MPLS TE tunnel. This example is only meant to enable an understanding of the proposed extension and does not illustrate every permutation of the MIB.

The following denotes the configured tunnel "head" entry:

```
In mplsTunnelTable:
{
mplsTunnelIndex              = 100,
mplsTunnelInstance           = 1,
mplsTunnelIngressLSRId       = 192.0.2.1,
mplsTunnelEgressLSRId        = 192.0.2.3,
mplsTunnelName               = "Tunnel",
...
mplsTunnelSignallingProto    = none (1),
mplsTunnelSetupPrio          = 0,
mplsTunnelHoldingPrio        = 0,
mplsTunnelSessionAttributes  = 0,
mplsTunnelLocalProtectInUse  = false (0),
mplsTunnelResourcePointer    = mplsTunnelResourceMaxRate.5,
mplsTunnelInstancePriority   = 1,
mplsTunnelHopTableIndex      = 1,
mplsTunnelIncludeAnyAffinity = 0,
mplsTunnelIncludeAllAffinity = 0,
mplsTunnelExcludeAnyAffinity = 0,
mplsTunnelPathInUse          = 1,
mplsTunnelRole               = head (1),
...
mplsTunnelRowStatus          = Active
}
```

BFD session parameters used to monitor this tunnel should be configured on head-end as follows:

```
In bfdSessTable:
BfdSessEntry ::= SEQUENCE {
-- BFD session index
bfdSessIndex                  = 2,
bfdSessVersionNumber          = 1,
-- LSP Ping used for OOB bootstrapping
bfdSessType  = multiHopOutOfBandSignaling,
...
bfdSessAdminStatus            = start,
...
```

```
        bfdSessDemandModeDesiredFlag  = false,
        bfdSessControlPlaneIndepFlag  = false,
        bfdSessMultipointFlag         = false,
        bfdSessDesiredMinTxInterval   = 100000,
        bfdSessReqMinRxInterval       = 100000,
        ...
        -- Indicates that the BFD session is to monitor an
           MPLS TE tunnel
        bfdSessExtMapType         = teIpv4(3),

        -- OID of the first accessible object (mplsTunnelName) of
           the mplsTunnelEntry identifying the MPLS TE tunnel (being
           monitored using BFD) in the MPLS  tunnel table.
           A value of zeroDotzero indicates that no association
           has been made as yet between the BFD session and the path
           being monitored.
           In the above OID example:
                100 -> Tunnel Index
                1 -> Tunnel instance
                3221225985 -> Ingress LSR Id 192.0.2.1
                3221225987 -> Egress LSR Id 192.0.2.3
        bfdSessExtMapPointer
                    = mplsTunnelName.100.1.3221225985.3221225987,
        bfdSessRowStatus      = createAndGo
       }
```

Similarly BFD session would be configured on the tail-end of the tunnel. Creating the above row will trigger the bootstrapping of the session using LSP Ping and its subsequent establishment over the path by de-multiplexing of the control packets using the BFD session discriminators.

5.3. BFD objects for session performance counters

BFD-STD-MIB defines BFD Session Performance Table (bfdSessionPerfTable), for collecting per-session BFD performance counters, as an extension to the bfdSessionTable.

The bfdSessionPerfTable is extended with the performance counters to collect Mis-connectivity Defect, Loss of Continuity Defect and RDI (Remote Defect Indication) counters.

1. bfdSessExtPerfMisDefCount - Mis-connectivity defect count for this BFD session.
2. bfdSessExtPerfLocDefCount - Loss of continuity defect count for this BFD session.
3. bfdSessExtPerfRdiInCount - Total number of RDI messages received for this BFD session.

4. bfdSessExtPerfRdiOutCount - Total number of RDI messages sent
for this BFD session.

5.4. Notification Objects

To be added in the next version of this document.

6. BFD MPLS-MPLS-TP MIB Module Definition

```
BFD-EXT-STD-MIB DEFINITIONS ::= BEGIN

IMPORTS
  MODULE-IDENTITY, OBJECT-TYPE, Counter32
      FROM SNMPv2-SMI             -- [RFC2578]

  RowPointer,TruthValue,TEXTUAL-CONVENTION
      FROM SNMPv2-TC              -- [RFC2579]

  MODULE-COMPLIANCE, OBJECT-GROUP
      FROM SNMPv2-CONF            -- [RFC2580]

  bfdSessIndex
      FROM BFD-STD-MIB;

  bfdExtMib MODULE-IDENTITY
  LAST-UPDATED "201110250000Z" -- October 25 2011
  ORGANIZATION "IETF Bidirectional Forwarding Detection
                Working Group"
  CONTACT-INFO
     "
        Sam Aldrin
        Huawei Technologies
        2330 Central Express Way,
        Santa Clara, CA 95051, USA
  Email:  aldrin.ietf@gmail.com

        Venkatesan Mahalingam
        Aricent
        India
  Email: venkat.mahalingams@gmail.com

        Mukund Mani
        Aricent



        India
  Email: mukund.mani@gmail.com
```

Kannan KV Sampath
Aricent
India
Email: Kannan.Sampath@aricent.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 is an initial version containing objects
              to provide a proactive mechanism to detect faults using
              BFD for MPLS and MPLS-TP networks"
        REVISION "201110250000Z" -- October 25 2011
        DESCRIPTION
            " Initial version published as RFC xxx "
        -- RFC Ed.: RFC-editor pls fill in xxxx
        ::=  { mib-2 XXX }  -- XXX to be replaced with correct value
        -- RFC Ed.: assigned by IANA

    -- ------------------------------------------------------------
    -- groups in the MIB
    -- ------------------------------------------------------------

          bfdExtObjects         OBJECT IDENTIFIER ::= { bfdExtMib 0 }
          bfdExtConformance     OBJECT IDENTIFIER ::= { bfdExtMib 1 }

    -- ------------------------------------------------------------
    -- Textual Conventions
    -- ------------------------------------------------------------

        SessionMapTypeTC ::= TEXTUAL-CONVENTION
            STATUS         current
            DESCRIPTION
              "Used to indicate the type of MPLS or MPLS-TP path
               associated to the session"
            SYNTAX INTEGER {
                     nonTeIpv4(1),    -- mapping into LDP IPv4
                     nonTeIpv6(2),    -- mapping into LDP IPv6
                     teIpv4(3),       -- mapping into TE IPv4
                     teIpv6(4),       -- mapping into TE IPv6
                     pw(5),           -- mapping into Pseudowires
```

```
                  mep(6)          -- mapping into MEPs in MPLS-TP
                }


   -- ---------------------------------------------------------------
   -- BFD session table extensions for BFD on MPLS and MPLS-TP
   -- ---------------------------------------------------------------
   -- bfdSessExtTable - bfdSessTable Extension

       bfdSessExtTable   OBJECT-TYPE
           SYNTAX              SEQUENCE OF BfdSessExtEntry
           MAX-ACCESS          not-accessible
           STATUS              current
           DESCRIPTION
             "This table is an extension to the bfdSessTable for
              configuring BFD sessions for MPLS or MPLS-TP paths."
       ::= { bfdExtObjects 1 }

       bfdSessExtEntry   OBJECT-TYPE
           SYNTAX              BfdSessExtEntry
           MAX-ACCESS          not-accessible
           STATUS              current
           DESCRIPTION
             "A row in this table extends a row in bfdSessTable."
         INDEX { bfdSessIndex }
       ::= { bfdSessExtTable 1 }

       BfdSessExtEntry ::= SEQUENCE {
           bfdSessExtRole               INTEGER,
           bfdSessExtMode               INTEGER,
           bfdSessExtTmrNegotiate       TruthValue,
           bfdSessExtMapType            SessionMapTypeTC,
           bfdSessExtMapPointer         RowPointer
       }

       bfdSessExtRole  OBJECT-TYPE
           SYNTAX      INTEGER {
                         active(1),
                         passive(2)
                       }
           MAX-ACCESS  read-create
           STATUS      current
           DESCRIPTION
             "This object specifies whether the system is playing the
              active(1) role or the passive(2) role for this
              BFD session."
           REFERENCE
               "RFC 5880, Section 6.1"
```

```
        DEFVAL { active }
    ::= { bfdSessExtEntry 1 }

    bfdSessExtMode  OBJECT-TYPE
        SYNTAX      INTEGER {
                      cc(1),
                      cv(2)
                    }
        MAX-ACCESS  read-create
        STATUS      current
        DESCRIPTION
          "This object specifies whether the BFD session is running
           in Continuity Check(CC) or the Connectivity
           Verification(CV) mode."
        REFERENCE
            "1.Proactive Connectivity Verification, Continuity Check
               and Remote Defect Indication for MPLS Transport
               Profile, draft-ietf-mpls-tp-cc-cv-rdi-06"
        DEFVAL { cc }
    ::= { bfdSessExtEntry 2 }

    bfdSessExtTmrNegotiate  OBJECT-TYPE
        SYNTAX             TruthValue
        MAX-ACCESS         read-create
        STATUS             current
        DESCRIPTION
          "This object specifies if timer negotiation is required for
           the BFD session. When set to false, timer negotiation is
           disabled"
        DEFVAL { true }
    ::= { bfdSessExtEntry 3 }

    bfdSessExtMapType OBJECT-TYPE
        SYNTAX             SessionMapTypeTC
        MAX-ACCESS         read-create
        STATUS             current
        DESCRIPTION
          "This object indicates the type of path being monitored
           by this BFD session entry."
    ::= { bfdSessExtEntry 4 }

    bfdSessExtMapPointer OBJECT-TYPE
        SYNTAX           RowPointer
        MAX-ACCESS       read-create
        STATUS           current
        DESCRIPTION
          "If bfdSessExtMapType is nonTeIpv4(1) or nonTeIpv6(2), then
           this object MUST contain zeroDotZero or point to
```

```
                an instance of the mplsXCEntry indicating the LDP-based
                LSP associated with this BFD session.



                If bfdSessExtMapType is teIpv4(3) or teIpv6(4), then
                this object MUST contain zeroDotZero or point to
                an instance of the mplsTunnelEntry indicating
                the RSVP-based MPLS TE tunnel associated with this
                BFD session.

                If bfdSessExtMapType is pw(5), then this object MUST
                contain zeroDotZero or point to an instance of
                the pwEntry indicating the MPLS Pseudowire associated
                with this BFD session.

                If bfdExtSessMapTpye is mep(6). then this object MUST
                contain zeroDotZero or point to an instance identifying
                the mplsOamIdMeEntry configured for monitoring the MPLS-TP
                path associated with this BFD session.

                If this object points to a conceptual row instance
                in a table consistent with bfdSessExtMapType but this
                instance does not currently exist then no valid
                path is associated with this session entry.

                If this object contains zeroDotZero then no valid path is
                associated with this BFD session entry till it is
                populated with a valid pointer consistent with
                the value of bfdSessExtMapType as explained above."
          ::= { bfdSessExtEntry 5 }

   -- ---------------------------------------------------------------
   -- BFD Objects for Session performance
   -- ---------------------------------------------------------------
   -- bfdSessExtPerfTable - bfdSessPerfTable Extension

       bfdSessExtPerfTable     OBJECT-TYPE
           SYNTAX              SEQUENCE OF BfdSessExtPerfEntry
           MAX-ACCESS          not-accessible
           STATUS              current
           DESCRIPTION
             "This table is an extension to the bfdSessPerfTable"
       ::= { bfdExtObjects 2 }

       bfdSessExtPerfEntry OBJECT-TYPE
           SYNTAX     BfdSessExtPerfEntry
           MAX-ACCESS not-accessible
```

```
            STATUS     current
            DESCRIPTION
                "A row in this table extends the bfdSessPerfTable"
         INDEX { bfdSessIndex }
        ::= { bfdSessExtPerfTable 1 }



        BfdSessExtPerfEntry ::= SEQUENCE {
            bfdSessExtPerfMisDefCount    Counter32,
            bfdSessExtPerfLocDefCount    Counter32,
            bfdSessExtPerfRdiInCount     Counter32,
            bfdSessExtPerfRdiOutCount    Counter32
        }

        bfdSessExtPerfMisDefCount OBJECT-TYPE
            SYNTAX          Counter32
            MAX-ACCESS      read-only
            STATUS          current
            DESCRIPTION
               "This object gives a count of the mis-connectivity defects
               detected for the BFD session. For instance, this count
               will be incremented when the received BFD control packet
               carries an incorrect globally unique source
               MEP identifier."
        ::= { bfdSessExtPerfEntry 1 }

        bfdSessExtPerfLocDefCount OBJECT-TYPE
            SYNTAX          Counter32
            MAX-ACCESS      read-only
            STATUS          current
            DESCRIPTION
               "This object gives a count of the Loss of continuity
                defects detected in MPLS and MPLS-TP paths"
        ::= { bfdSessExtPerfEntry 2 }

        bfdSessExtPerfRdiInCount OBJECT-TYPE
            SYNTAX          Counter32
            MAX-ACCESS      read-only
            STATUS          current
            DESCRIPTION
               "This object gives a count of the Remote Defect
                Indications received for the BFD session."
        ::= { bfdSessExtPerfEntry 3 }

        bfdSessExtPerfRdiOutCount OBJECT-TYPE
            SYNTAX          Counter32
            MAX-ACCESS      read-only
```

```
            STATUS          current
            DESCRIPTION
               "This object gives a count of the Remote Defect
                Indications sent by the BFD session"
        ::= { bfdSessExtPerfEntry 4 }

        -- Module compliance



        bfdExtGroups
        OBJECT IDENTIFIER ::= { bfdExtConformance 1 }

        bfdExtCompliances
        OBJECT IDENTIFIER ::= { bfdExtConformance 2 }

        -- Compliance requirement for fully compliant implementations.

        bfdExtModuleFullCompliance MODULE-COMPLIANCE
        STATUS current
        DESCRIPTION
        "Compliance statement for agents that provide full
        support for the BFD-EXT-STD-MIB module. "

        MODULE -- This module.

        MANDATORY-GROUPS {
            bfdSessionExtGroup,
            bfdSessionExtPerfGroup
        }
        ::= { bfdExtCompliances 1 }

        bfdExtModuleReadOnlyCompliance MODULE-COMPLIANCE
        STATUS current
        DESCRIPTION
        "Compliance requirement for implementations that only
        provide read-only support for BFD-EXT-STD-MIB. Such devices
        can then be monitored but cannot be configured using
        this MIB module."

        MODULE -- This module.

        MANDATORY-GROUPS {
            bfdSessionExtGroup,
            bfdSessionExtPerfGroup
        }
```

```
        OBJECT       bfdSessExtRole
        MIN-ACCESS   read-only
        DESCRIPTION  "Write access is not required."

        OBJECT       bfdSessExtMode
        MIN-ACCESS   read-only
        DESCRIPTION  "Write access is not required."

        OBJECT       bfdSessExtTmrNegotiate
        MIN-ACCESS   read-only
        DESCRIPTION  "Write access is not required."



        OBJECT       bfdSessExtMapType
        MIN-ACCESS   read-only
        DESCRIPTION  "Write access is not required."

        OBJECT       bfdSessExtMapPointer
        MIN-ACCESS   read-only
        DESCRIPTION  "Write access is not required."

        ::= { bfdExtCompliances 2 }


        -- Units of conformance.

       bfdSessionExtGroup OBJECT-GROUP
       OBJECTS {
                    bfdSessExtRole,
                    bfdSessExtMode,
                    bfdSessExtTmrNegotiate,
                    bfdSessExtMapType,
                    bfdSessExtMapPointer
             }
       STATUS     current
       DESCRIPTION
             "Collection of objects needed for BFD monitoring for
             MPLS and MPLS-TP paths"
             ::= { bfdExtGroups 1 }

       bfdSessionExtPerfGroup OBJECT-GROUP
       OBJECTS {
                   bfdSessExtPerfMisDefCount,
                   bfdSessExtPerfLocDefCount,
                   bfdSessExtPerfRdiInCount,
                   bfdSessExtPerfRdiOutCount
             }
```

```
    STATUS     current
    DESCRIPTION
          "Collection of objects needed to monitor the
           performance of BFD sessions on MPLS and MPLS-TP
           paths"
          ::= { bfdExtGroups 2 }


    END
```

7. Security Considerations

To be added in the next version of this document.

8. IANA Considerations

To be added in the next version of this document.

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.

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

[BFD-1HOP] Katz, D. and D. Ward, "Bidirectional Forwarding Detection (BFD) for IPv4 and IPv6 (Single Hop)", RFC 5881, June 2010.

[BFD-MH] Katz, D. and D. Ward, "Bidirectional Forwarding Detection (BFD) for Multihop Paths", RFC 5883, June 2010.

[BFD-MPLS] Aggarwal, R. et.al., "Bidirectional Forwarding Detection (BFD) for MPLS Label Switched Paths (LSPs)", RFC 5884, June 2010

[MPLS-TP-CC-CV-RDI] Allan, D., G. Swallow, Drake, J., "Proactive Connectivity Verification, Continuity Check and Remote Defect Indication for MPLS Transport Profile", draft-ietf-mpls-tp-cc-cv-rdi-06.

[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.

9.2 Informative References

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

10. Acknowledgments

The authors would like to thank Lavanya Srivatsa for her valuable comments.

11. Authors' Addresses

Sam Aldrin
Huawei Technologies
2330 Central Express Way,
Santa Clara, CA 95051, USA
Email: aldrin.ietf@gmail.com

Venkatesan Mahalingam
Aricent
India
Email: venkat.mahalingams@gmail.com

Mukund Mani
Aricent
India
Email: mukund.mani@gmail.com

Kannan KV Sampath
Aricent
India
Email: Kannan.Sampath@aricent.com

Thomas D. Nadeau
CA Technologies
273 Corporate Drive, Portsmouth, NH, USA
Email: thomas.nadeau@ca.com

Network Working Group
INTERNET-DRAFT
Intended Status: Standards Track
Expires: April 19, 2012

Sam Aldrin
Huawei Technologies
M.Venkatesan
Kannan KV Sampath
Aricent Group
Thomas D. Nadeau
CA Technologies
Sami Boutros
Cisco Systems
Ping Pan
Infinera

October 17, 2011

MPLS-TP Operations, Administration, and Management (OAM) Identifiers Management Information Base (MIB)
draft-vkst-mpls-tp-oam-id-mib-01

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/1id-abstracts.txt.

The list of Internet-Draft Shadow Directories can be accessed at

http://www.ietf.org/shadow.html.

This Internet-Draft will expire on April 19, 2012.

Copyright and License Notice

Table of Contents

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, Pseudowires and Sections.

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 [RFC6370].

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 [RFC6370]. 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 [RFC6370] (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, RowPointer,
          DisplayString, StorageType
             FROM SNMPv2-TC                    -- [RFC2579]
          mplsStdMIB
             FROM MPLS-TC-STD-MIB              -- [RFC3811]
          InterfaceIndexOrZero, ifGeneralInformationGroup,
          ifCounterDiscontinuityGroup
             FROM IF-MIB;                      -- [RFC2863]

      mplsOamIdStdMIB MODULE-IDENTITY
         LAST-UPDATED
            "201110170000Z" -- October 17, 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: venkat.mahalingams@gmail.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
          "201110170000Z" -- October 17, 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 MplsOamIdMegEntry
        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        MplsOamIdMegEntry
        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 [RFC6370] (MEG_ID for LSP is LSP_ID and
               MEG_ID for PW is PW_Path_ID), MEP_ID can be
               automatically formed.

               For co-routed bidirectional LSP, MEG_ID is
               A1-{Global_ID::Node_ID::Tunnel_Num}::Z9-{Global_ID::
               Node_ID::Tunnel_Num}::LSP_Num.

               For associated bidirectional LSP, MEG_ID is A1-
               {Global_ID::Node_ID::Tunnel_Num::LSP_Num}:: Z9-
```

```
                {Global_ID::Node_ID::Tunnel_Num::LSP_Num}

                For LSP, MEP_ID is formed using,
                Global_ID::Node_ID::Tunnel_Num::LSP_Num

                For PW, MEG_ID is formed using AGI::A1-
                {Global_ID::Node_ID::AC_ID}:: Z9-
                {Global_ID::Node_ID::AC_ID}.

                For PW, MEP_ID is formed using
                AGI::Global_ID::Node_ID::AC_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. RFC 6371, Operations, Administration, and Maintenance
                Framework for MPLS-Based Transport Networks,
                September 2011.
             3. RFC 6370, MPLS Transport Profile (MPLS-TP) Identifiers.
             4. MPLS-TP Identifiers Following ITU-T Conventions
                [TP-ITUIDS]."
        INDEX { mplsOamIdMegIndex }
        ::= { mplsOamIdMegTable 1 }

        MplsOamIdMegEntry ::= SEQUENCE {
             mplsOamIdMegIndex            Unsigned32,
             mplsOamIdMegName             DisplayString,
             mplsOamIdMegOperatorType     INTEGER,
             mplsOamIdMegIdIcc            DisplayString,
             mplsOamIdMegIdUmc            DisplayString,
             mplsOamIdMegServiceType      INTEGER,
             mplsOamIdMegMpLocation       INTEGER,
             mplsOamIdMegRowStatus        RowStatus,
             mplsOamIdMegStorageType      StorageType
        }

        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
            "1. RFC 6370, MPLS Transport Profile (MPLS-TP)
                Identifiers.
             2. MPLS-TP Identifiers Following ITU-T Conventions
                [TP-ITUIDS]."
        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 MplsOamIdMegOperatorType value is iccBased(2),
           otherwise null ICC value should be assigned."
        REFERENCE
```

```
            "MPLS-TP Identifiers Following ITU-T Conventions
             [TP-ITUIDS]."
         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 MplsOamIdMegOperatorType value is iccBased(2),
            otherwise null ICC value should be assigned."
         REFERENCE
            "MPLS-TP Identifiers Following ITU-T Conventions
           [TP-ITUIDS]."
         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
          "RFC 6371, Operations, Administration, and Maintenance
           Framework for MPLS-Based Transport Networks,
           September 2011."
        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 }

    mplsOamIdMegStorageType 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 }
         ::= { mplsOamIdMegEntry 9 }

    -- End of MPLS Transport Profile MEG table
```

```
       -- Start of MPLS Transport Profile ME table
       mplsOamIdMeTable OBJECT-TYPE
        SYNTAX        SEQUENCE OF MplsOamIdMeEntry
        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        MplsOamIdMeEntry
        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 }

       MplsOamIdMeEntry ::= SEQUENCE {
            mplsOamIdMeIndex                  Unsigned32,
            mplsOamIdMeMpIndex                Unsigned32,
            mplsOamIdMeName                   DisplayString,
            mplsOamIdMeMpIfIndex              InterfaceIndexOrZero,
            mplsOamIdMeSourceMepIndex         Unsigned32,
            mplsOamIdMeSinkMepIndex           Unsigned32,
            mplsOamIdMeMpType                 INTEGER,
            mplsOamIdMeMepDirection           INTEGER,
            mplsOamIdMeProactiveOamSessIndex  Unsigned32,
            mplsOamIdMeProactiveOamPhbTCValue INTEGER,
            mplsOamIdMeOnDemandOamPhbTCValue  INTEGER,
            mplsOamIdMeServicePointer         RowPointer,
            mplsOamIdMeRowStatus              RowStatus,
            mplsOamIdMeStorageType            StorageType
       }

       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, used to create
             multiple MEPs in a node of single ME. 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
            "RFC 6371, Operations, Administration, and Maintenance
             Framework for MPLS-Based Transport Networks,
             September 2011."
          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 {
                          ef1 (1),
                          ef2 (2),
                          af1 (3),
                          af2 (4),
                          af3 (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 { ef1 }
          ::= { mplsOamIdMeEntry 10 }

      mplsOamIdMeOnDemandOamPhbTCValue  OBJECT-TYPE
          SYNTAX        INTEGER {
                          ef1 (1),
                          ef2 (2),
                          af1 (3),
                          af2 (4),
                          af3 (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 { ef1 }

          ::= { mplsOamIdMeEntry 11 }

       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
             poinit-to-multipoint, where each ME will point to different
             branches of a P2MP tree."
          DEFVAL { zeroDotZero }
          ::= { mplsOamIdMeEntry 12 }

       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 13 }

       mplsOamIdMeStorageType 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 }
            ::= { 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,
         mplsOamIdMegStorageType
      }

      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,
         mplsOamIdMeServicePointer,
         mplsOamIdMeRowStatus,
         mplsOamIdMeStorageType
      }
      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.

[TP-ITUIDS] Winter, R., van Helvoort, H., and M. Betts, "MPLS-TP Identifiers Following ITU-T Conventions", ID draft-ietf-mpls-tp-itu-t-identifiers-00, July 2011.

[RFC6370] Bocci, M., Swallow, G., and E. Gray, "MPLS-TP Identifiers", RFC 6370, September 2011.

[RFC6371] Busi, I., Niven-Jenkins, B., and D. Allan, "MPLS-TP OAM Framework and Overview", RFC 6371, September 2011.

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: venkat.mahalingams@gmail.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

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: April 9, 2012

IJ. Wijnands, Ed.
Cisco Systems, Inc.
P. Hitchen
BT
N. Leymann
Deutsche Telekom
W. Henderickx
Alcatel-Lucent
October 7, 2011

Multipoint Label Distribution Protocol In-Band Signaling in a VPN Context draft-wijnands-mpls-mldp-vpn-in-band-signaling-00

Abstract

Sometimes an IP multicast distribution tree (MDT) traverses both MPLS-enabled and non-MPLS-enabled regions of a network. Typically the MDT begins and ends in non-MPLS regions, but travels through an MPLS region. In such cases, it can be useful to begin building the MDT as a pure IP MDT, then convert it to an MPLS Multipoint LSP (Label Switched Path) when it enters an MPLS-enabled region, and then convert it back to a pure IP MDT when it enters a non-MPLS-enabled region. [I-D.ietf-mpls-mldp-in-band-signaling] specifies the procedures for building such a hybrid MDT, using Protocol Independent Multicast (PIM) as the control protocol in the non-MPLS region of the network, and using Multipoint Extensions to Label Distribution Protocol (mLDP) in the MPLS region. This document extends those procedures so that they will work when the links between the MPLS and non-MPLS regions are [RFC4364] interfaces. While these procedures do not provide a good general multicast VPN solution, they are useful in certain specific situations.

Status of this Memo

This Internet-Draft will expire on April 9, 2012.

Copyright Notice

Table of Contents

1. Introduction

Sometimes an IP multicast distribution tree (MDT) traverses both MPLS-enabled and non-MPLS-enabled regions of a network. In such cases, it can be useful to begin building the MDT as a pure IP MDT, then convert it to an MPLS Multipoint LSP (Label Switched Path) when it enters an MPLS-enabled region, and then convert it back to a pure IP MDT when it enters a non-MPLS-enabled region. [I-D.ietf-mpls-mldp-in-band-signaling] specifies the procedures for building such a hybrid MDT, using Protocol Independent Multicast (PIM) [RFC4601] as the control protocol in the non-MPLS region of the network, and using Multipoint Extensions to Label Distribution Protocol (mLDP) [I-D.ietf-mpls-ldp-p2mp] in the MPLS region.

For a given tree, one or more routers that are at the border between a non-MPLS-enabled region and an MPLS-enabled region receive PIM control messages from the non-MPLS-enabled region, and convert them to mLDP control messages to be sent into the MPLS-enabled region. Another router that is at the border between a non-MPLS-enabled region and an MPLS-enabled region receives mLDP control messages from the MPLS-enabled region and converts them to PIM control messages to be sent into a non-MPLS-enabled region.

In PIM, a tree is identified by a source address (or in the case of bidirectional trees [RFC5015], a rendezvous point address or "RPA") and a group address. The tree is built from the leaves up, by sending PIM control messages in the direction of the source address or the RPA. In mLDP, a tree is identified by a root address and an "opaque value", and is built by sending mLDP control messages in the direction of the root. The procedures of [I-D.ietf-mpls-mldp-in-band-signaling] explain how to convert a PIM <source address or RPA, group address> pair into an mLDP <root node, opaque value> pair;, and how to convert the mLDP <root node, opaque value> pair back into the original PIM <source address or RPA, group address> pair.

However, those procedures assume that the routers doing the PIM/mLDP conversion have routes to the source address or RPA in their global routing tables. Thus the procedures cannot be applied exactly as specified when the interfaces connecting the non-MPLS-enabled region to the MPLS-enabled region are interfaces that belong to a VPN as described in [RFC4364]. This specification extends the procedures of [I-D.ietf-mpls-mldp-in-band-signaling] so that they may be applied in the VPN context.

As in [I-D.ietf-mpls-mldp-in-band-signaling], the scope of this document is limited to the case where the multicast content is distributed in the non-MPLS-enabled regions via PIM-created Source-

Specific or Bidirectional trees. Bidirectional trees are always mapped onto Multipoint-to-Multipoint LSPs, and source-specific trees are always mapped onto Point-to-Multipoint LSPs.

Each multicast tree in the non-MPLS enabled region is mapped 1-1 onto a multipoint LSP in the MPLS-enabled region. For a variety of reasons (discussed in [I-D.ietf-l3vpn-2547bis-mcast]), this is not a suitable for a general purpose multicast VPN solution. But the procedures described herein are much simpler than the general purpose MVPN procedures, and are applicable when the 1-1 mapping is acceptable, and when it is acceptable to use mLDP as the protocol for setting up the multipoint LSPs. For example, some service providers offer multicast content to their customers, but have chosen to use VPNs to isolate the various customers and services. This is a very different scenario than the general MVPN scenario, in which the customers provide their own multicast content, out of the control of the service provider.

Due to the 1-1 mapping and the multicast source and group information being encoded in the mLDP FEC, there is deterministic mapping beween the multicast tree and the mLDP LSP in the core network. This improves and simplifies fault resolution.

In order to use the mLDP in-band signaling procedures for a particular group address in a particular VPN, the Provider Edge (PE) routers that attach to that VPN MUST be configured with a range of multicast group addresses for which mLDP in-band signaling is to be enabled. This configuration is per VRF ("Virtual Routing and Forwarding table", defined in [RFC4364]). For those groups, and those groups only, the procedures of this document are used instead of the general purpose Multicast VPN procedures. This configuration must be present in all PE routers that attach to sites containing senders or receivers for the given set of group addresses.

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 RFC 2119 [RFC2119].

1.2. Terminology

IP multicast tree : An IP multicast distribution tree identified by an source IP address and/or IP multicast destination address, also referred to as (S,G) and (*,G).

mLDP : Multicast LDP.

In-band signaling : Using the opaque value of a mLDP FEC element to encode the (S,G) or (*,G) identifying a particular IP multicast tree.

P2MP LSP: An LSP that has one Ingress LSR and one or more Egress LSRs (see [I-D.ietf-mpls-ldp-p2mp]).

MP2MP LSP: An LSP that connects a set of leaf nodes, acting indifferently as ingress or egress (see [I-D.ietf-mpls-ldp-p2mp]).

MP LSP: A multipoint LSP, either a P2MP or an MP2MP LSP.

Ingress LSR: Source of a P2MP LSP, also referred to as root node.

2. VPN In-band signaling for MP LSPs

Suppose that a PE router, PE1, attaching to a particular VPN, receives a PIM Join(S,G) message over one of its VRF interfaces. Following the procedure of section 5.1 of [I-D.ietf-l3vpn-2547bis-mcast], PE1 determines the "upstream RD", the "upstream PE", and the "upstream multicast hop" (UMH) for the source address S. Please note that sections 5.1.1 and 5.1.2 of [I-D.ietf-l3vpn-2547bis-mcast] are applicable.

In order to transport the multicast tree via a MP LSP using VPN in-band signaling, an mLDP Label Mapping Message is sent by PE1. This message will contain either a P2MP FEC or an MP2MP FEC (see [I-D.ietf-mpls-ldp-p2mp], depending upon whether the PIM tree being transported is a source-specific tree, or a bidirectional tree, respectively. The FEC contains a "root" and an "opaque value".

If the UMH and the upstream PE have the same IP address (i.e., the Upstream PE is the UMH), then the root of the Multipoint FEC is set to the IP address of the Upstream PE. If, in the context of this VPN, (S,G) refers to a source-specific MDT, then the values of S, G, and the upstream RD are encoded into the opaque value. If, in the context of this VPN, G is a bidirectional group address, then S is replaced with the value of the RPA associated with G. The coding details are specified in Section 3. Conceptually, the Multipoint FEC can be thought of as an ordered pair: <root=Upstream-PE,

opaque_value=<S or RPA ,G ,Upstream-RD>. The mLDP Label Mapping Message is then sent by PE1 on its LDP session to the "next hop" on its path to the upstream PE. The "next hop" is usually the IGP next hop, but see [I-D.napierala-mpls-targeted-mldp] for cases in which the next hop is not the IGP next hop.

If the UMH and the upstream PE do not have the same IP address, the procedures of section 2 of [I-D.ietf-mpls-mldp-recurs-fec] should be applied. The root node of the multipoint FEC is set to the UMH. The recursive opaque value is then set as follows: the root node is set to the upstream PE, and the opaque value is set to the multipoint FEC described in the previous paragraph. That is, the multipoint FEC can be thought of as the following recursive ordered pair: <root=UMH, opaque_value=<root=Upstream-PE, opaque_value =<S or RPA, G, Upstream-RD>>.

The encoding of the multipoint FEC also specifies the "type" of PIM MDT being spliced onto the multipoint LSP. Four types of MDT are defined: IPv4 source-specific tree, IPv6 source-specific tree, IPv4 bidirectional tree, and IPv6 bidirectional tree.

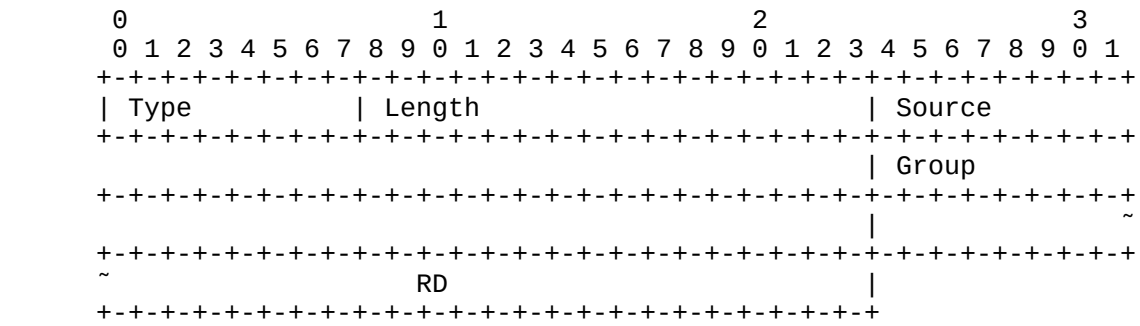
When a PE router receives an mLDP message with a P2MP or MP2MP FEC, where the PE router itself is the root node, and the opaque value is one of the types defined in Section 3, then it uses the RD encoded in the opaque value field to determine the VRF context. (This RD will be associated with one of the PEs VRFs.) Then, in the context of that VRF, the PE follows the procedure specified in section 2 of [I-D.ietf-mpls-mldp-in-band-signaling].

3. Encoding the Opaque Value of an LDP MP FEC

This section documents the different transit opaque encodings.

3.1. Transit VPNv4 Source TLV

This opaque value type is used when transporting a source-specific mode multicast tree whose source and group addresses are IPv4 addresses.



Type: (to be assigned by IANA).

Length: 16

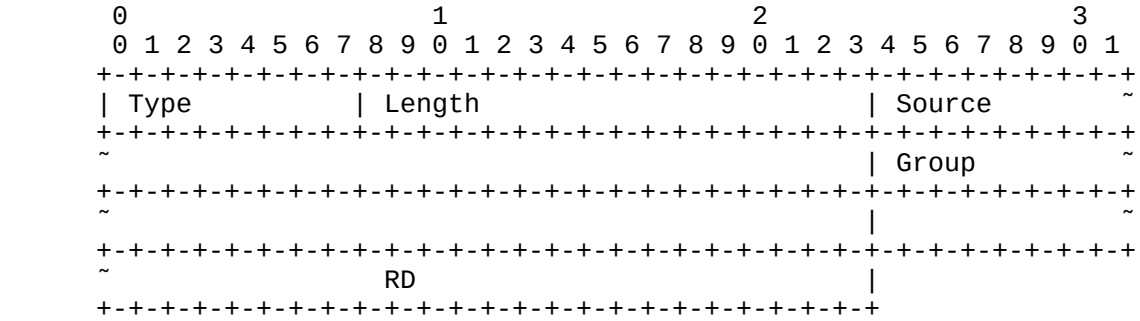
Source: IPv4 multicast source address, 4 octets.

Group: IPv4 multicast group address, 4 octets.

RD: Route Distinguisher, 8 octets.

3.2. Transit VPNv6 Source TLV

This opaque value type is used when transporting a source-specific mode multicast tree whose source and group addresses are IPv6 addresses.



Type: (to be assigned by IANA).

Length: 40

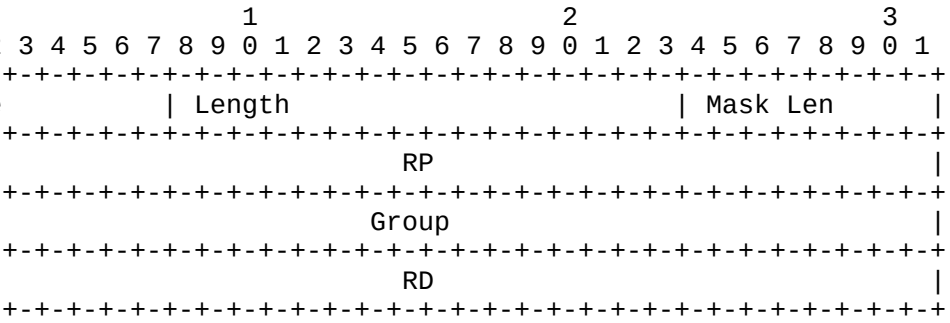
Source: IPv6 multicast source address, 16 octets.

Group: IPv6 multicast group address, 16 octets.

RD: Route Distinguisher, 8 octets.

3.3. Transit VPNv4 bidir TLV

This opaque value type is used when transporting a bidirectional multicast tree whose group address is an IPv4 address. The RP address is also an IPv4 address in this case.



Type: (to be assigned by IANA).

Length: 17

Mask Len: The number of contiguous one bits that are left justified and used as a mask, 1 octet.

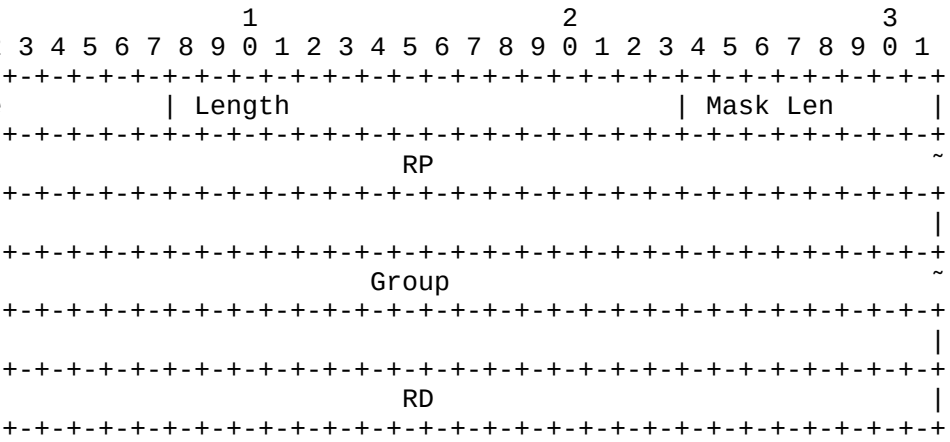
RP: Rendezvous Point (RP) IPv4 address used for encoded Group, 4 octets.

Group: IPv4 multicast group address, 4 octets.

RD: Route Distinguisher, 8 octets.

3.4. Transit VPNv6 bidir TLV

This opaque value type is used when transporting a bidirectional multicast tree whose group address is an IPv6 address. The RP address is also an IPv6 address in this case.



Type: (to be assigned by IANA).

Length: 41

Mask Len: The number of contiguous one bits that are left justified and used as a mask, 1 octet.

RP: Rendezvous Point (RP) IPv6 address used for encoded group, 16 octets.

Group: IPv6 multicast group address, 16 octets.

RD: Route Distinguisher, 8 octets.

4. Security Considerations

The same security considerations apply as for the base LDP specification, described in [RFC5036], and the base mLDP specification, described in [I-D.ietf-mpls-ldp-p2mp]

5. IANA considerations

[I-D.ietf-mpls-ldp-p2mp] defines a registry for "The LDP MP Opaque Value Element Basic Type". This document requires the assignment of four new code points in this registry:

Transit VPNv4 Source TLV type - requested 10

Transit VPNv6 Source TLV type - requested 11

Transit VPNv4 Bidir TLV type - requested 12

Transit VPNv6 Bidir TLV type - requested 13

6. Acknowledgments

Thanks to Eric Rosen and Andy Green for their comments on the draft.

7. References

7.1. Normative References

[RFC4364] Rosen, E. and Y. Rekhter, "BGP/MPLS IP Virtual Private Networks (VPNs)", RFC 4364, February 2006.

[RFC4601] Fenner, B., Handley, M., Holbrook, H., and I. Kouvelas, "Protocol Independent Multicast - Sparse Mode (PIM-SM): Protocol Specification (Revised)", RFC 4601, August 2006.

[RFC5015] Handley, M., Kouvelas, I., Speakman, T., and L. Vicisano, "Bidirectional Protocol Independent Multicast (BIDIR-PIM)", RFC 5015, October 2007.

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

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

[I-D.ietf-mpls-ldp-p2mp] Minei, I., Kompella, K., Wijnands, I., and B. Thomas, "Label Distribution Protocol Extensions for Point-to-Multipoint and Multipoint-to-Multipoint Label Switched Paths", draft-ietf-mpls-ldp-p2mp-11 (work in progress), October 2010.

[I-D.ietf-mpls-mldp-in-band-signaling] Wijnands, I., Eckert, T., Leymann, N., and M. Napierala, "mLDP based in-band signaling for Point-to-Multipoint and Multipoint-to- Multipoint Label Switched Paths", draft-ietf-mpls-mldp-in-band-signaling-02 (work in progress), July 2010.

[I-D.ietf-mpls-mldp-recurs-fec] Wijnands, I., Rosen, E., Napierala, M., and N. Leymann, "Using mLDP through a Backbone where there is no Route to the Root", draft-ietf-mpls-mldp-recurs-fec-00 (work in progress), October 2010.

[I-D.ietf-l3vpn-2547bis-mcast] Aggarwal, R., Bandi, S., Cai, Y., Morin, T., Rekhter, Y., Rosen, E., Wijnands, I., and S. Yasukawa, "Multicast in MPLS/BGP IP VPNs", draft-ietf-l3vpn-2547bis-mcast-10 (work in progress), January 2010.

7.2. Informative References

[I-D.napierala-mpls-targeted-mldp] Napierala, M., Rosen, E., and I. Wijnands, "Using LDP Multipoint Extensions on Targeted LDP Sessions", draft-napierala-mpls-targeted-mldp-00 (work in progress), January 2011.

Authors' Addresses

IJsbrand Wijnands (editor)
Cisco Systems, Inc.
De kleetlaan 6a
Diegem 1831
Belgium

Email: ice@cisco.com

Paul Hitchen
BT
BT Adastral Park
Ipswich IP53RE
UK

Email: paul.hitchen@bt.com

Nicolai Leymann
Deutsche Telekom
Winterfeldtstrasse 21
Berlin 10781
Germany

Email: n.leymann@telekom.de

Wim Henderickx
Alcatel-Lucent
Copernicuslaan 50
Antwerp 2018
Belgium

Email: wim.henderickx@alcatel-lucent.com

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: April 26, 2012

Q. Zhao
E. Chen
Huawei Technology
October 24, 2011

Protection Mechanisms for Label Distribution Protocol P2MP/MP2MP Label Switched Paths
draft-zhao-mpls-mldp-protections-00.txt

Abstract

Service providers continue to deploy real-time multicast applications using Multicast LDP (mLDP) across MPLS networks. There is a clear need to protect these real-time applications and to provide the shortest switching times in the event of failure. This document outlines the requirements, describes the protection mechanisms available, and where neccessary proposes extensions to facilitate mLDP P2MP and MP2MP LSP protection within an MPLS network.

Table of Contents

1. Terminology

For a clear narrative, this section gives a general conceptional overview of the terms.

- PLR: The node where the traffic is logically redirected onto the preset backup path is called Point of Local Repair.

- MP: The node where the backup path merges with the primary path is called Merge Point.

- FD: The node that detects the failure on primary path, and then triggers the action(s) for traffic protection is called Failure Detector. Either traffic sender or receiver can be the FD, depending on which protection mode are deployed. More details are described in later sections of this document.

- SP: The node where the traffic is physically switched/duplicated onto the backup path is called Switchover Point. In multicast cases, PLR and SP can be two different nodes. More details are described in later sections of this document.

2. 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 RFC 2119 [RFC2119].

3. Introduction

In order to meet user demands, operators and service providers continue to deploy multicast applications using mLDP across MPLS networks. In certain scenarios, traditional IGP-mLDP convergence mechanisms fail to meet protection switching times required to minimise, or negate entirely, application interruptions for real-time applications, including stock trading, on-line games, and multimedia teleconferencing.

Current best practice for protecting services, and higher applications includes the pre-computation and establishment of a backup path, this can decrease the convergence time efficiently. Once a failure has been detected on the primary path, the traffic should be transmitted across the back-up path.

However, two major challenges exist with the aforementioned solution. The first is how to build an absolutely disjointed backup path for

each node in a multicast tree; the second is how to balance between convergence time and resource consumption.

This document provides several ways to setup the backup path for mLDP LSP, including local protection, territorial protection, and end-to-end protection. The goal is to build a reliable umbrella to against traffic black hole. How to detect failure is outside the scope of this document.

More and more users are apt to deploy multicast applications on MPLS mLDP network. In some scenarios, traditional IGP-mLDP convergence is hard to meet the requirements of those real-time applications, such as stock business, on-line game, and multimedia teleconference.

The industry has reached a consensus that setting up a backup path previously can decrease the convergence time efficiently. No matter how the above-mentioned backup path was established, once the failure is detected, the traffic should be transmitted at that path as soon as possible. Even so, there are still two major challenges left for us, one is how to build an absolutely disjointed backup path for each node in a multicast tree; the other is how to balance between convergence time and resource consumption.

It is getting urgent to find the ideal protection mechanism(s) to improve the convergence time, and at the meantime minimize the side-effects, such as bandwidth wastage.

For a primary LDP P2MP/MP2MP LSP, there are several ways to set up its backup path. It can use RSVP-TE P2P tunnel as a logical out-going interface, consequently utilize the mature high availability technologies of RSVP-TE. Or, it can make use of LDP P2P backup LSP as a packet encapsulation, so that the complex configuration of P2P RSVP-TE can be skipped. Or, it can build its own P2MP/MP2MP backup LSP according to IGP’s loop-free alternative route, thus avoid double label stack. Other than these, it can also build a totally disjointed LSP in another topology, accordingly take advantage of the real end-to-end protection.

When the backup path is present, there are two options for packet forwarding and switchover. If the traffic sender feeds the stream on both paths, and the traffic receiver drops packet on backup path, the switchover will be very quick once the failure is detected, because the whole switchover action is a local behavior on traffic receiver. The disadvantage of this manner is that traffic will be duplicated on both paths, and consume double bandwidth. Contrastively, if the traffic sender feeds stream only on the primary path, the resource wastage can be waived. Cooperation is needed in this manner, so there will be some protocol extensions. But if the performance can

be equal or better than the previous option, it is reasonable to choose the second one.

This document describes several methods to setup and switch paths for options to setup the backup LDP P2MP/MP2MP LSP. mLDP LSPs, including local protection, territorial protection, and end-to-end protection. The goal is to identify strengths, weaknesses and gaps, in order to build a reliable set of tools to shield against traffic black holes that would severely impact real-time applications, in the event of primary path failure.

3.1. Requirements

A number of requirements have been identified that allow the optimal set of mechanisms to developed. These currently include:

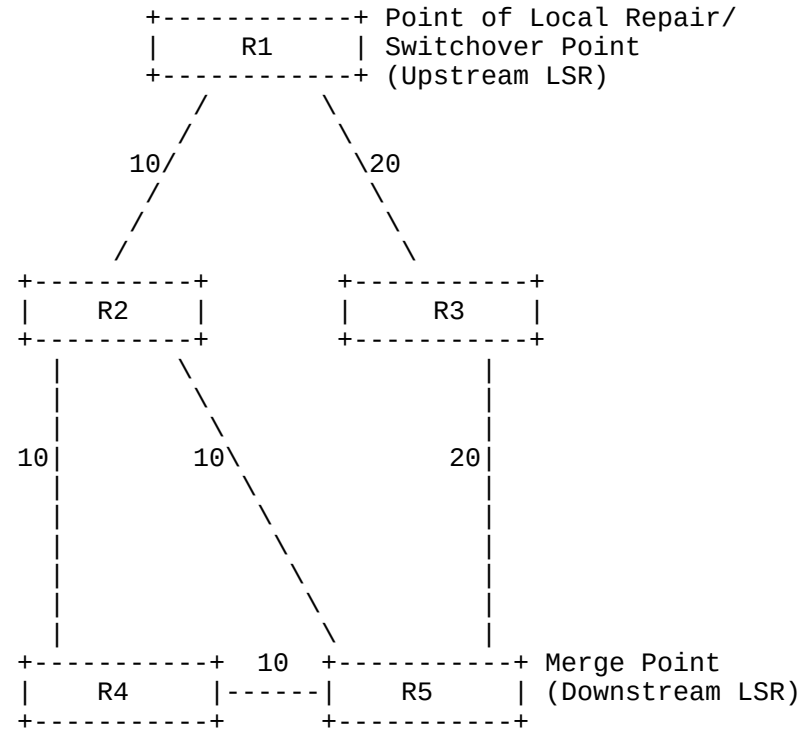
- Computation of a disjointed (link and node) backup path within the multicast tree;
- Minimisation of protection convergence time;
- Optimisation of bandwitdth usage.

3.2. Scope

The method to detect failure is outside the scope of this document. Also this document does not provide any authorization mechanism for controlling the set of LSRs that may attempt to join a mLDP protection session.

4. Local protection using P2P LSP

By encapsulating mLDP packets within an P2P TE tunnel or P2P LDP backup LSP, the LDP P2MP/MP2MP LSP can be protected by the P2P protection mechanisms. However, this protection mechanism is not capable of detecting, and recovering, if the failure occurs on the destination node of the P2P backup LSP. Thus, this section provides a unified method to protect both node and link with P2P backup LSP.



mLDP Local Protection Example

Figure 1

In Figure 1 (mLDP Local Protection Example) above, the preferential path from R1 to R4/R5 is through R2, and the secondary path is through R3. In this case, the mLDP LSP will be established according to the IGP preferential path as R1--R2--R4/R5.

It is the responsibility of R2 to inform R1 of its downstream LSRs (in this example R4 and R5) and the respective labels (L4 and L5). Once the link between R1 and R2 fails, or R2 node fails, R1 will duplicate the traffic to R4 and R5, with inner label as L4/L5, and outer label as the P2P backup LSP R1--R3--R5--R4 and R1--R3--R5.

Finally, the previous forwarding states will be removed after R4 and R5 finish their Make-Before-Break (MBB) procedure.

4.1. Signaling procedures for local protection

Continuing to use Figure 1 (mLDP Local Protection Example), R2 sends a notification message to R1 to inform the node that R2 has two downstream nodes, R4 and R5 with forwarding labels L4 and L5 respectively.

When R1 sees R2 node going down, it takes mLDP packets as it would send them to R4 and R5 through R2 and sends them into the two P2P backup tunnels:

- P2P tunnel R1--R3--R5--R4, using inner label L4.

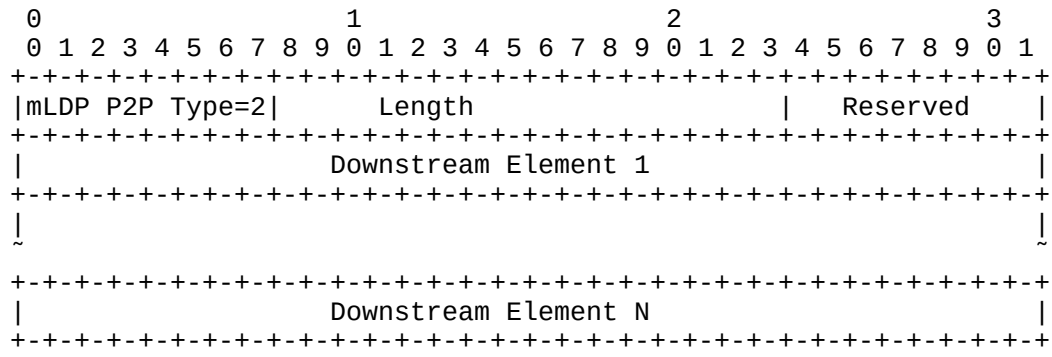
- P2P tunnel R1--R3--R5, using inner label L5.

So that R4/R5 will receive same packets as from the interface between R2 and R4/R5, just from different interface.

At the same time, R1 sends notifications with MBB request status code to R4 and R5. So that after R4 and R5 are done with MBB, they will send the notifications to R3 with MBB done status code. And then R3 will remove the old forwarding state which is being protected by the P2P backup tunnels.

4.2. Protocol extensions for local protection

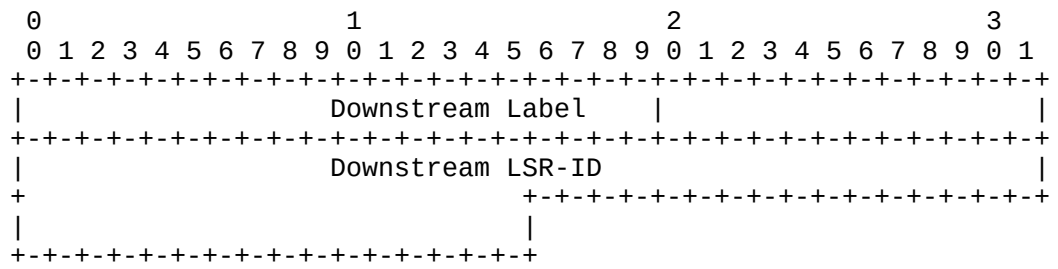
A new type of LDP MP Status Value Element is introduced, for notifying downstream LSRs and respective labels. It is encoded as follows:



mLDP P2P Encapsulation Status Code

Figure 2

The Downstream Element is encoded as follows:

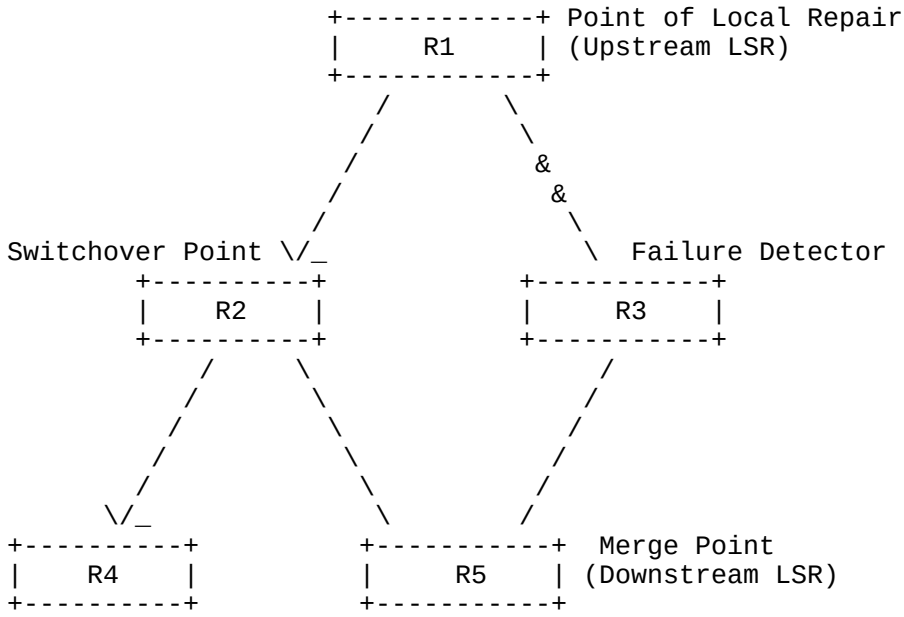


Downstream Element in mLDP P2P Encapsulation Status Code

Figure 3

5. Territorial protection using mLDP LSP

Making use of IGP-FRR results, LDP can build the backup mLDP LSP for territorial protection. Note that in some scenarios, such as the following example, Failure Detector and Point of Local Repair, Switchover Point and Merge Point can be different nodes.



mLDP Territorial Protection Example

Figure 4

In Figure 4 (mLDP Territorial Protection Example), normally R1 feeds traffic to R4 through R2, and feeds traffic to R5 through R3. Once the link between R1 and R3 fails, R1 will be the logical Point of Local Repair node, which feeds the traffic to R5 through backup path on R2. Because R2 is already receiving traffic, so that R1 does not need to take any action. It is responsibility of R2 to duplicate the traffic to R5, as a Switchover Point. In this case, as the Failure Detector, R3 will need to send out the notification to R2, in order to trigger the switchover procedure.

5.1. Signaling Procedures for Territorial Protection

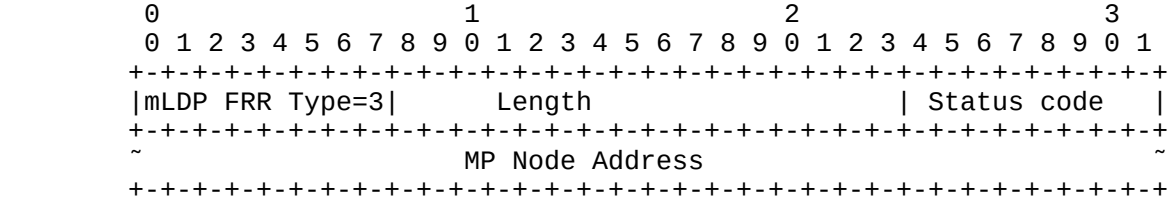
Merge Point (R5) determines the primary and secondary paths according to the IGP-FRR results. Then it sends out label mapping message including an LDP MP Status TLV that carries a FRR Status Code to indicate the primary path and secondary path. At the same time, it triggers a reverse path for failure notification by sending out label request message with an LDP MP Status TLV. The reverse path is uniquely identified by root address, opaque value, and MP address.

When failure is detected by Failure Detector (R3), it will send out the failure notification, then traffic will switch to the secondary path.

When Merge Point (R5) sees the next hop to Root changed, it will advertise the new mapping, and the traffic will re-converge to the new primary path.

5.2. Protocol extensions for Territorial Protection

A new type of LDP MP Status Value Element is introduced, for setting up secondary mLDP LSP. It is encoded as follows:



mLDP FRR Status Code

Figure 5

mLDP FRR Type: Type 3 (to be assigned by IANA)

Length: If the Address Family is IPv4, the Address Length MUST be 5; if the Address Family is IPv6, the Address Length MUST be 17.

Status code: 1 = Primary path for traffic forwarding (used in Label Mapping message)
2 = Secondary path for traffic forwarding (used in Label Mapping message)
3 = Reverse path for failure notification (used in Label Request message)
4 = Failure notification (used in Notification message)

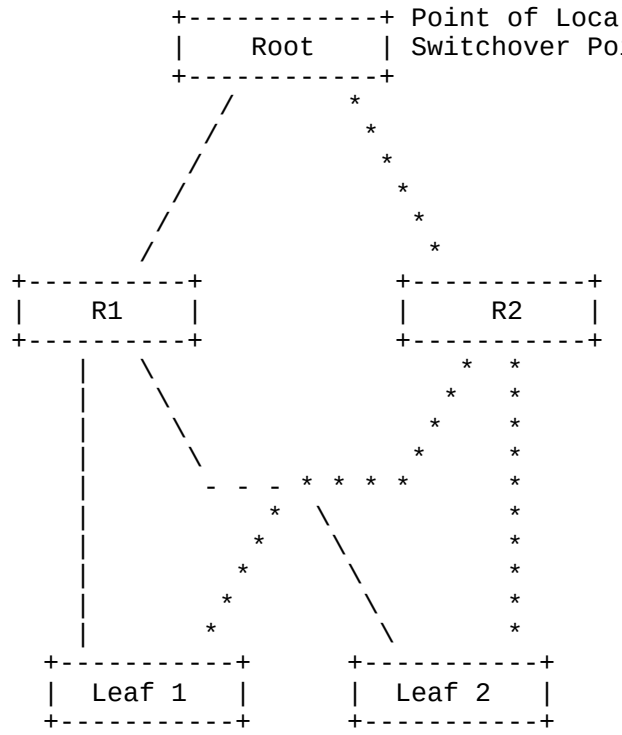
MP Node Address: A host address encoded according to the Address Family of this LSP.

mLDP Bandwidth Reservation Status Code Parameters

Figure 6

6. End-to-end protection using LDP Multiple Topology

[I-D.ietf-mpls-ldp-multi-topology] provides a mechanism to setup disjointed LSPs within different topologies. So that applications can use these redundant LSPs for end-to-end protection.



mLDP End-to-end Protection Example

Figure 7

In Figure 7 (mLDP End-to-end Protection Example), there are two separated topologies from Root node to Leaf 1 and Leaf 2. For the same FEC element, the Leaf node can trigger mLDP LSPs in each topology. Root node can setup 1:1 or 1+1 end-to-end protection, using these two mLDP LSPs.

6.1. Signaling Procedures for End-to-end Protection

Using Figure 7 (mLDP Local Protection Example), Leaf 1 and Leaf 2 may trigger mLDP LSPs in different topologies, sending label mapping

messages with same FEC element, different MT-ID and different label. When the Root node receives the label mapping messages from different topologies, it will set up two mLDP LSPs for application as end-to-end protection. Failure detection for the primary mLDP LSP is outside the scope of this document. But either Root node or Leaf node can be the Failure Detector.

6.2. Protocol extensions for End-to-end Protection

The protocol extensions required to build mLDP LSPs in different topologies is defined in [I-D.ietf-mpls-ldp-multi-topology].

7. Acknowledgements

We would like to thank authors of draft-ietf-mpls-mp-ldp-reqs and the authors of draft-ietf-mpls-ldp-multi-topology from which some text of this document has been inspired.

8. IANA Considerations

This memo includes the following requests to IANA:

- mLDP P2P Encapsulation type for LDP MP Status Value Element.
- mLDP FRR type for LDP MP Status Value Element.

9. Manageability Considerations

[Editors Note - This section requires further discussion]

9.1. Control of Function and Policy

9.2. Information and Data Models

9.3. Liveness Detection and Monitoring

9.4. Verifying Correct Operation

9.5. Requirements on Other Protocols and Functional Component

9.6. Impact on Network Operation

9.7. Policy Control

10. Security Considerations

The same security considerations apply as for the base LDP specification, as described in [RFC5036]. The protocol extensions specified in this document do not provide any authorization mechanism for controlling the set of LSRs that may attempt to join a mLDP protection session. If such authorization is desirable, additional mechanisms, outside the scope of this document, are needed.

Note that authorization policies should be implemented and/or configure at all the nodes involved .

11. References

11.1. Normative References

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

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

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

[RFC5561] Thomas, B., Raza, K., Aggarwal, S., Aggarwal, R., and JL. Le Roux, "LDP Capabilities", RFC 5561, July 2009.

[RFC6348] Le Roux, JL. and T. Morin, "Requirements for Point-to-Multipoint Extensions to the Label Distribution Protocol", RFC 6348, September 2011.

[I-D.ietf-mpls-ldp-p2mp] Minei, I., Wijnands, I., Kompella, K., and B. Thomas, "Label Distribution Protocol Extensions for Point-to-Multipoint and Multipoint-to-Multipoint Label Switched Paths", draft-ietf-mpls-ldp-p2mp-15 (work in progress), August 2011.

[I-D.ietf-mpls-ldp-multi-topology] Zhao, Q., Fang, L., Zhou, C., Li, L., So, N., and R. Torvi, "LDP Extension for Multi Topology Support", draft-ietf-mpls-ldp-multi-topology-00 (work in progress), October 2011.

11.2. Informative References

[RFC3468] Andersson, L. and G. Swallow, "The Multiprotocol Label Switching (MPLS) Working Group decision on MPLS signaling protocols", RFC 3468, February 2003.

Authors' Addresses

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

Email: quintin.zhao@huawei.com

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

Email: emily.chenying@huawei.com

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: April 25, 2012

R. Zheng
L. Jin
ZTE

October 23, 2011

LSP Ping extensions for echo reply relay
draft-zj-mpls-lsp-ping-reply-relay-00

Abstract

[RFC4379] describes the LSP Ping mechanism to detect data plane failures. In some deployment scenario for the LSP traceroute, a replying LSR may not have the available route to the initiator, and the echo reply message sent to the initiator would be discarded. Thus, the basic idea of traceroute procedure to localize fault could not be achieved. This document describes extensions to LSP Ping mechanism to enable the replying LSR to have the capability to relay the echo reply by a set of routable intermediate nodes to the initiator.

Status of this Memo

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

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

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

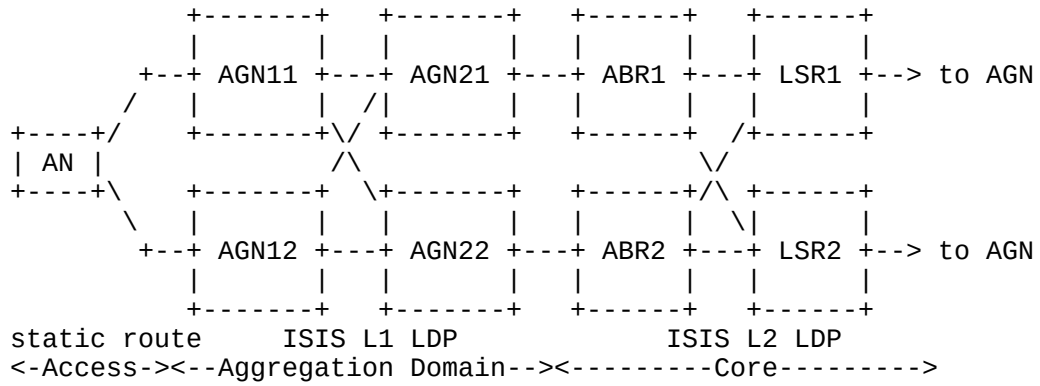
This Internet-Draft will expire on April 25, 2012.

Table of Contents

1. Introduction

LSP Ping is an efficient OAM mechanism to detect data plane failures and localize faults. The basic LSP Ping mechanism has been described in [RFC4379]. In traceroute mode of LSP Ping procedure, the echo request message is sent to the control plane of each transit LSR, and an echo reply massage with proper information are required to send to the initiator at each transit LSR. Then the LSP fault could be localized exactly, and an accurate LSP topology could also be built.

The echo reply would normally be sent back to the initiator via an IPv4/IPv6 UDP packet. The basic requirement is that the replying LSR has reachable IP route to the initiator. However, in some network deployment, the requirement could not be met because of the route control policy. For example, considering the inter-area situation in Seamless MPLS architecture [ietf-mpls-seamless], LSR1/LSR2 nodes in core network would not have IP reachable route to ANs. If tracing an LSP from AN to remote AN, the LSR1/LSR2 node would be unable to respond to the echo request message for the lack of IP reachable route to AN.



This draft describes extensions to LSP Ping mechanism to enable the echo reply message to be relayed back to the initiator. The replying LSR would send the echo reply message to a routable relayed node indicated by the Routable Relay Node Address TLV, and the echo reply would be relayed to the next relay node, till to the initiator.

Note: [I-D. ietf-mpls-interas-lspping-00] describes an echo reply relayed mechanism for inter-AS scenario, by using a dedicated ASBR stack list. Unfortunately, this mechanism could not be applied in

inter-area scenario. The current mechanism described in this draft is only for inter-area scenario.

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. Routable Relay Node Address TLV

The Routable Relay Node Address TLV MUST be carried in the echo request and echo reply messages if the echo reply relayed mechanism described in this draft is required.

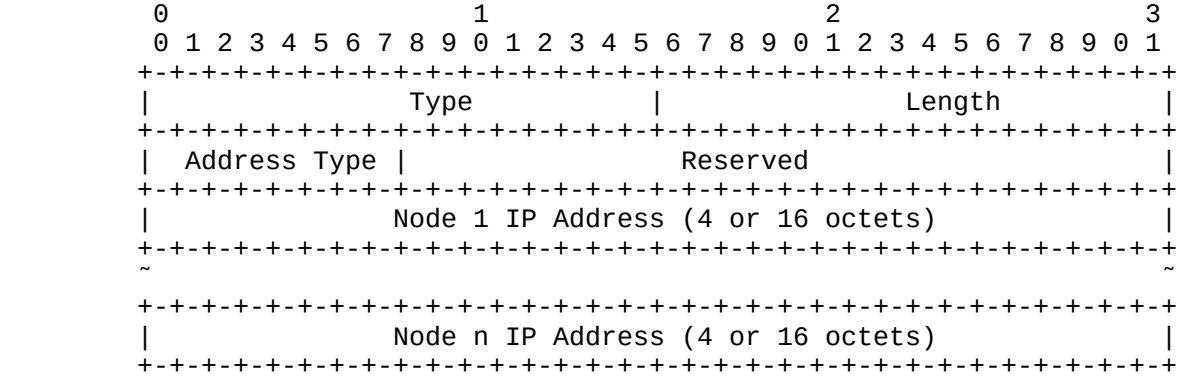


Figure 1: Routable Relay Node Address TLV

- Type: to be assigned by IANA.

- Length: The Length of the Value field in octets.

- Address Type

The Address Type indicates the routable relay node address type. The routable relay node IP address length is determined based on the Address Type.

Type	Address Type	Node Address length
1	IPv4	4
2	IPv6	16

- Node 1 IP Address: The IP Address of the first node that adds its address in this TLV.

- Node n IP Address: The IP Address of the last node that adds its address in this TLV.

3. Procedures

When tracing an LSP with outmost label TTL=1, a Routable Relay Node Address TLV with the first address item of the initiator address should be included in the echo request.

Upon receiving an echo request message, the receiver checks the address list in the Routable Relay Node Address TLV from node 1 to n IP address. Until finds the first routable node address, sets this node address as the destination address of the echo reply message. The receiver deletes all the node IP address items behind of the first routable one in the address list, and adds its own address as the last address item in the Routable Relay Node Address TLV. The updated Routable Relay Node Address TLV should be carried in the echo reply message.

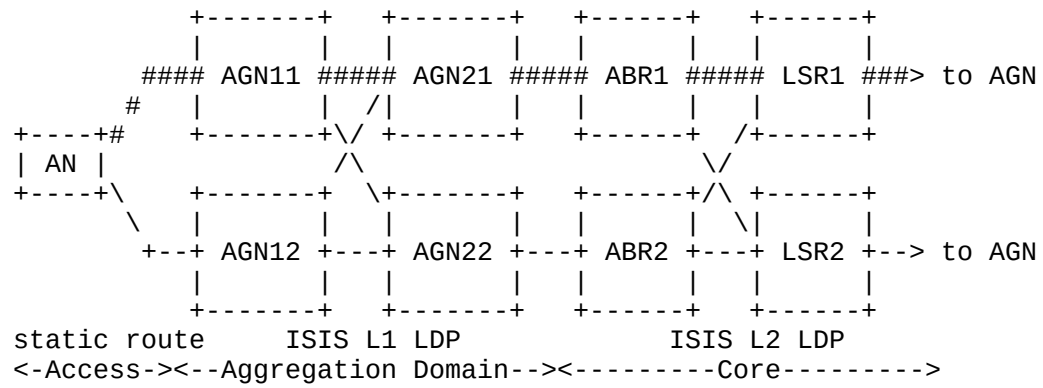
Upon receiving an echo reply message with its address as the destination address in the IP header, the receiver should check the address items in Routable Relay Node Address TLV in sequence to find the next relay node address. If its own address is node n IP address, then the next relay node address is node (n-1) IP address. The receiver should then send the echo reply to the next relay node with the Routable Relay Node Address Stack TLV unchanged. The next relay node may be an intermediate node or the initiator.

As relayed by intermediate nodes, the echo reply will finally be transmitted to the initiator. The initiator will copy the Routable Relay Node Address TLV from the echo reply into the next echo request message.

Those nodes do not understand or support the Routable Relay Node Address TLV, SHOULD ignore and pass it unchanged.

4. LSP Ping Example

Considering the inter-area scenario of Seamless MPLS architecture below,



Supposing the active LSP from AN to remote AN goes through AGN11, AGN21, ABR1, LSR1 and the following nodes. When performing LSP traceroute on the LSP the first echo request sent by AN with outmost label TTL=1, contains the Routable Relay Node Address TLV with the only address of AN.

After processed by AGN11, AGN11 address will be added in the Routable Relay Node Address list following AN1 address in the echo reply.

AN1 copies the Routable Relay Node Address TLV into the next echo request when receiving the echo reply.

Upon receiving the echo request, AGN21 checks the address list in the Routable Relay Node Address TLV in sequence, and finds out that AN1 address is routable. Then deletes AGN11 address, and adds its own address following AN1 address. As a result, there would be AN1 address followed by AGN21 address in the Routable Relay Node Address TLV of the echo reply sent by AGN21.

For echo request with outmost label TTL=3, similarly with the procedures on AGN21, the Routable Relay Node Address TLV sent by ABR1 in the echo reply will include two address items with both AN1 and ABR1 address.

Then AN1 sends echo request with outmost label TTL=4, containing the Routable Relay Node Address TLV copied from the received echo reply message. Upon receiving the echo request message, LSR1 checks the address list in the Routable Relay Node Address TLV in sequence, and finds out that AN1 address is IP route unreachable, and ABR1 address is the first routable one in the Routable Relay Node Address TLV. LSR1 adds its address as the last address item following ABR1 address in the Routable Relay Node Address TLV, sets ABR1 address as the

destination address of the echo reply, and sends the echo reply to ABR1.

Upon receiving the echo request message from LSR1, ABR1 checks the address list in the Routable Relay Node Address TLV in sequence, and finds out that AN1 address is the one just before its address. Then ABR1 sends the echo reply to AN1 with the Routable Relay Node Address TLV unchanged.

The echo reply from the replying node which has no reachable route to the initiator is finally transmitted to the initiator by multiple relay nodes.

5. Security Considerations

TBD

6. IANA Considerations

TBD

7. References

7.1. Normative References

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

[RFC4379] Kompella, K. and G. Swallow, "Detecting Multi-Protocol Label Switched (MPLS) Data Plane Failures", RFC 4379, February 2006.

7.2. Informative References

[I-D.ietf-mpls-interas-lspping-00]
Nadeau, T.and G. Swallow, "Detecting MPLS Data Plane Failures in Inter-AS and inter-provider Scenarios", draft-ietf-mpls-interas-lspping-00(expired) , March 2007.

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

Authors and contributors' Addresses

Ryan Zheng
ZTE Corporation
50, Ruanjian Avenue
Nanjing, 210012, China

Email: zheng.zhi@zte.com.cn

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

Email: lizhong.jin@zte.com.cn

Manuel Paul
Deutsche Telekom
Goslarer Ufer 35
10589 Berlin, Germany

Email: manuel.paul@telekom.de