

MPLS Working Group
Internet-Draft
Intended Status: Experimental RFC
Expires: February 2013

Bhargav Bhikkaji
Balaji Venkat Venkataswami
DELL-Force10
Shankar Raman
Gaurav Raina
I.I.T Madras
August 25, 2012

An Architecture for splicing TE-LSPs in Hierarchical CsC scenarios
draft-balaji-mpls-csc-te-lsp-splice-01

Abstract

Hierarchical Carrier Supporting Carrier deployments involve a Carrier Core which hereinafter is called the Tier-1 provider and two or more VPN sites that are carriers themselves hereinafter called Tier-2 providers that offer MPLS VPN services to their own customers. In such cases normally LDP is used to distribute labels amongst the routers (P and PE devices) in the Tier-2 provider's sites. When RSVP based TE-LSPs are constructed to explicitly route traffic for Tier-2 ISP's customers from the Tier-2 PEs to the CE of the Tier-1 provider and such TE-LSPs exist on multiple sites of the Tier-2 provider, the Tier-2 ISP may require splicing together through an "auto-match-and-splice-together" facility such that traffic flows from the PE of the Tier-2 ISP through the TE-LSP onto the CE of the Tier-1 ISP and then onto the other site and takes a path through a specific TE-LSP from the CE of the other site to the destination Tier-2 PE and then onto the final customer.

This solution offers a lot of advantages such as providing adequate assurance that the bandwidth for the traffic flowing through these spliced TE-LSPs is met. It also provides a explicit routing of the traffic rather than through the regular LDP (which follows IGP) scenarios. Such explicitly routed TE-LSPs would have been constructed taking into account factors such as using under-utilized links for example. Splicing together these TE-LSPs in various sites and doing the splicing on an auto-match based on bandwidth or delay metrics would be a good service to offer to the Tier-2 ISPs customers.

This draft outlines a scheme that offers such a feature and service to the Tier-2 ISPs through the addition of certain additional label exchanges and some additional labels such as the RSVP-stitch label and the RSVP-splicing-LDP label in the label stack which can be used to splice together these tunnels.

Status of this Memo

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

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

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

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

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

Copyright and License Notice

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

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

Table of Contents

1	Introduction	4
1.1	Terminology	8
2.0	Constructing spliced TE-LSPs between the Tier-2 sites . . .	8
2.0.1	RSVP-splicing-LDP label	10
2.0.2	RFC 6511 applicability	10
2.1	Decision at CE-B or the upstream CE in the Tier-2 ISP site. .	10
2.1.1	RSVP-stitch label	10
2.2	Across the Carrier's Core	11
2.3	Decision at PE-Lo	11

2.4	Decision at CE-B	11
2.5	Multiplicity of TE-LSP sections	12
2.6	Illustration	12
2.7	Utility	15
3	Security Considerations	17
4	IANA Considerations	17
5	References	17
5.1	Normative References	17
5.2	Informative References	18
	Authors' Addresses	18

1 Introduction

Some ISP customers of the MPLS/VPN backbone may want to provide MPLS/VPN services for their own customers. These Tier-2 ISPs that we will call them henceforth obtain the services for MPLS/VPN from a Top level Carrier which we will henceforth call as a Tier-1 ISP for their connectivity when these Tier-2 ISPs do not have networks that span across the region, between geographical regions or across the globe.

This type of connectivity is known as hierarchical VPN sometimes referred to as recursive VPN. Its deployment is similar to the other Carrier's Carrier VPNs except Multi-protocol iBGP is introduced for the distribution of the prefixes and label information between ISP sites.

An example topology is provided below...

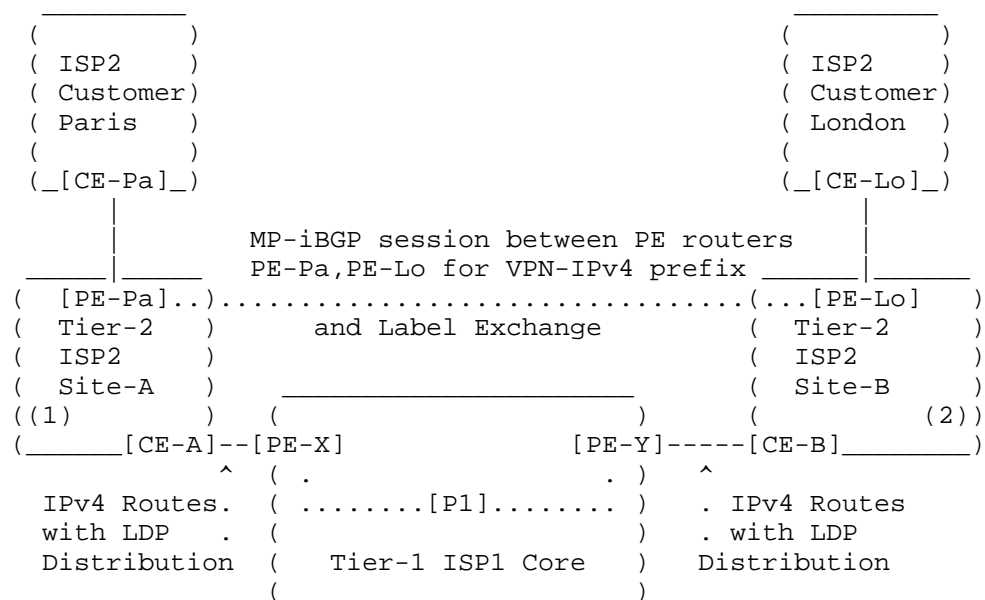


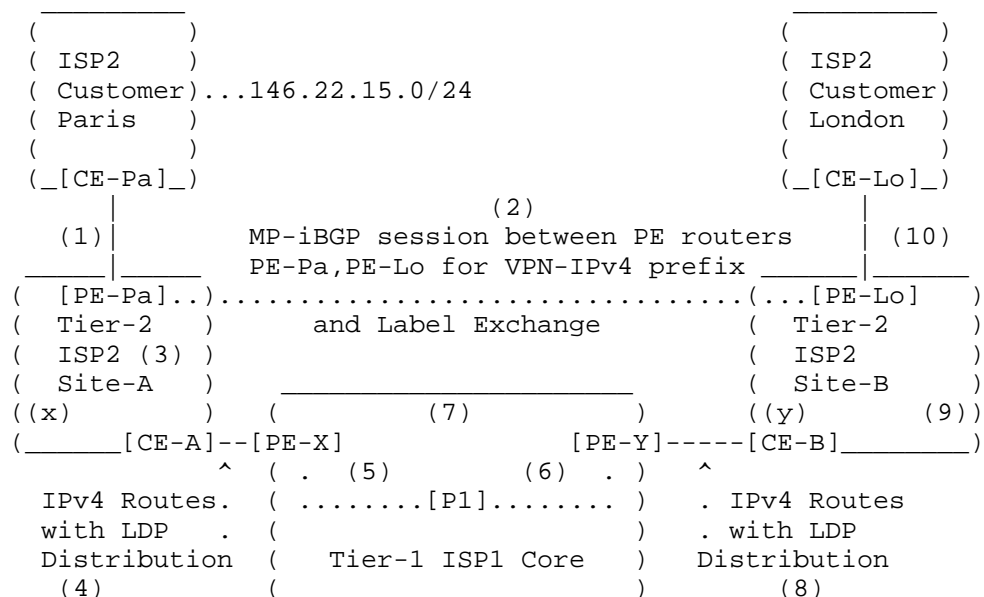
Figure 1: Reference Architecture Diagram

Within each Tier-2 ISP site in Figure 1 the PE-Routers PE-Pa and PE-Lo hold VRF routing information for any VPN customers that are attached to the POP at Paris and London. This is no different than

the standard MPLS/VPN architecture so VPN-IPv4 prefixes are assigned to customer VPN routes and are distributed between Tier-2 ISP sites using MP-iBGP with BGP extended community attributes (Router Target and Site of Origin).

Because each POP site for the Tier-2 ISP at London and Paris may hold several PE routers a full mesh of MP-iBGP is necessary among all PE-routers within the ISP MPLS/VPN topology. However again route reflectors can be deployed to cut down on the number of required peering sessions. In the example shown in Figure 1 it would be possible for example for the Tier-2 ISP2 London and Paris PE-routers to be route reflectors for their own Tier-2 ISP site. One could even deploy totally separate devices and make each PE-router a route reflector client so that MP-iBGP updates can be successfully reflected to all PE-routers that need the VPN information contained within the updates.

In the following figure 2 we provide an example of the relevant label assignment for the 146.22.15.0/24 prefix which is learned from a VPN customer of the Tier-2 ISP in the Paris area.



Legend :

(x) IGP routes with LDP distribution

(y) IGP routes with LDP distribution

Figure 2: Reference Diagram for normal control plane exchange for HCSc

Legend for the control plane exchange3 is as follows :

(1) CE-Pa sends update for 146.22.15.0/24 to PE-Pa

(2) An MP-iBGP update for Net=146.22.15.0/24 with next hop as PE-Pa and label assignment as 99 is sent to PE-Lo from PE-Pa

(3) An IGP + LDP update for Net=PE-Pa with label as pop action is sent to CE-A from PE-Pa

(4) The CE-A device sends an update with Net=PE-Pa with NH=CE-A and a label assignment of 1 to PE-X.

(7) An MP-iBGP update is sent from PE-X to PE-Y with Net=PE-Pa NH=PE-X and Label as 4.

(5) An LDP update goes from PE-X with Net as PE-X and Label as pop action to P1.

- (6) An LDP update goes from P1 with Net=PE-X and label as 2 to PE-Y
- (8) An LDP update with Net=PE-Pa and NH=PE-Y with label as 3 from PE-Y to CE-B.
- (9) An IGP update goes from CE-B with Net=PE-Pa with NH as PE-Y to PE-Lo
- (10) An LDP Update goes from CE-B to PE-Lo with Net=PE-Pa and label as 5.

The figure shows again that labels are advertised both within the MPLS/VPN backbone and within each ISP site, for each of the Tier-2 ISP (Tier-1's customer) internal routes. ISP-customer (Tier-2 customer) external routes are distributed between the sites as VPN-IPv4 routes. This mean that the iBGP session between sites becomes an MP-iBGP session so that the VPN-IPv4 routes and associated labels can be successfully advertised.

The actual traffic flow for a packet destined for a host on the 146.22.15.0/24 subnet and arriving at the Tier-2 ISP's PE-Lo router is illustrated in figure 2.

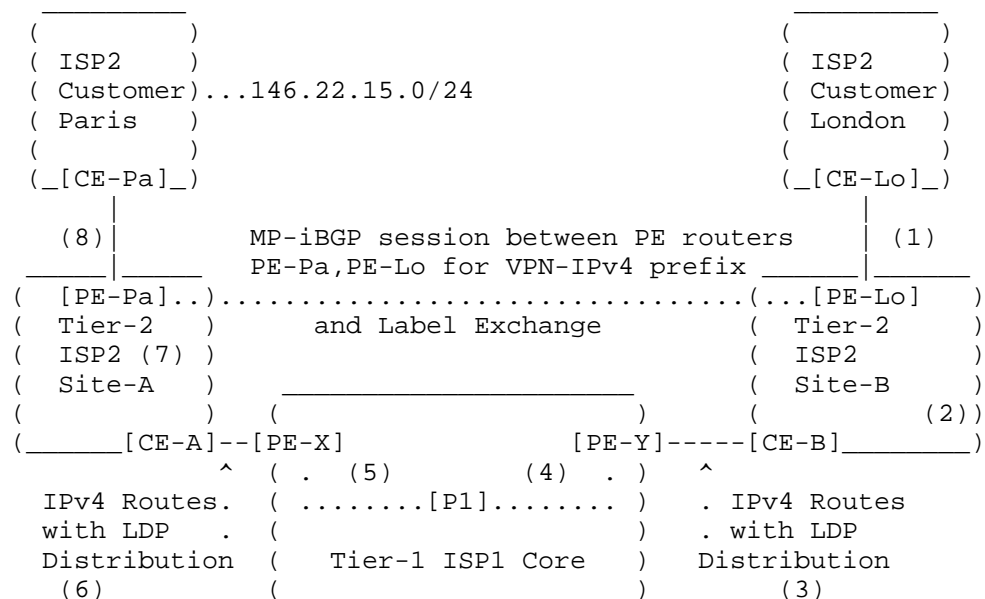


Figure 2: Reference Diagram for normal Data Plane exchange for HCSC

- (1) IP packet with IP-DA as 146.22.15.1
- (2) Label packet with MPLS labels 5,99, IP-DA 146.22.15.1
- (3) Label packet with MPLS labels 3,99, IP-DA 146.22.15.1
- (4) Label packet with MPLS labels 2,4,99, IP-DA 146.22.15.1
- (5) Label packet with MPLS labels 4,99, IP-DA 146.22.15.1
- (6) Label packet with MPLS labels 1,99, IP-DA 146.22.15.1
- (7) Label packet with MPLS labels 99, IP-DA 146.22.15.1
- (8) IP packet with IP-DA as 146.22.15.1

1.1 Terminology

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

2.0 Constructing spliced TE-LSPs between the Tier-2 sites

It is possible that there be requirements to establish TE LSPs for a

variety of reasons between the PE routers in the Tier-2 sites for example between PE-Lo and PE-Pa. These variety of reasons could include Traffic engineering necessities that may arise for the sake of allocating a certain amount of bandwidth for sets of traffic from the Tier-2 customer sites or for guaranteeing a certain delay budget or merely for utilizing under-utilized links (which otherwise would not have been used if left to IGP paths).

Consider a situation where there exists requirements to establish TE-LSPs between PE-Lo and PE-Pa for traffic direction from PE-Lo towards PE-Pa.

These TE-LSP needs to traverse the Tier-1 ISP core. To traverse the core it is possible to envisage that there exists sufficient bandwidth between PE-X and PE-Y. One possibility is to establish a TE-LSP between PE-X and PE-Y and use that LSP as a forwarding adjacency LSP for carrying the traffic over the core. The other possibility is to use normal LDP to carry the traffic over the core.

In both cases the TE-LSP established at the Tier-2 Customer sites need to be spliced together. And that splicing should be done automatically with reference to the characteristics that the TE-LSP advertises. In case of using both schemes for traversing the core, the TE-LSP in Site-B needs to be spliced to the section in Site-A.

Again it is possible that the section of the TE-LSP in Site-B was constructed independently of Site-A TE-LSP section. The TE-LSP section in Site-B being between PE-Lo and CE-B and the Site-A TE-LSP section between CE-A and PE-Pa.

Now there has to be a mechanism of conveying the section information between Site-A and Site-B. For traffic direction from Site-B to Site-A the draft solution that we propose intends to convey this TE-LSP information with TE-LSP characteristics such as bandwidth, delay, cost et... through a MP-iBGP update from CE-A to CE-B. This mechanism conveys the existence of a TE-LSP between PE-Pa and CE-A.

For the reverse direction of traffic the MP-iBGP update for the vice-versa occurs.

In our case in the diagram Figure 3 the PE-Pa-to-CE-A TE-LSP is advertised to CE-B. This information is thus sent from CE-A to CE-B. This is sent thus as a MP-iBGP update. This MP-iBGP update is sent to the PE-Pa and the PE-Lo Provider Edge routers as well. This is required since the PE-Lo in our example can take a decision to use the TE-LSP or the normal LDP path at its end based on knobs configured or based on certain policy decisions at that time of sending the traffic where such policies could be configured. These

policy decisions could be built in as an algorithm with a set of rules. This mechanism is outside the scope of the current document.

2.0.1 RSVP-splicing-LDP label

CE-B then generates two different labels towards PE-Y one for LDP and another for RSVP-splicing-LDP. The LDP label is used when the packet reaches CE-B towards PE-Y when the labels in the packet are LDP based labels. If the packet arrives with a RSVP Label for the TE-LSP between PE-Lo and CE-B at the head of the stack of labels at CE-B then the RSVP-splicing-LDP label is used.

This also means that the MP-iBGP exchange between PE-X and PE-Y has to have two classes of labels one for LDP and another for RSVP-splicing-LDP.

Additionally PE-X to CE-A labels have to be of two kinds. One for LDP and another for RSVP-splicing-LDP.

2.0.2 RFC 6511 applicability

For RSVP tunnels proposed in this mechanism it is important that non-PHP behaviour be observed by the egress LSRs in the Carrier's core and in the TE-LSP sections in the Tier-2 ISP as per [RFC6511].

2.1 Decision at CE-B or the upstream CE in the Tier-2 ISP site.

If the CE-B in our example has received MP-iBGP updates about the TE-LSP at the remote site (CE-A to PE-Lo) it can take a decision as to whether it has to use that TE-LSP or use an ordinary LDP based LSP by choosing either the LDP label or the RSVP-splicing-LDP label. MP-iBGP updates can be expected to keep the information of the TE-LSPs at the remote Tier-2 site current by periodically but not too often updating the information through a MP-iBGP update. This MP-iBGP update should have characteristics of the TE-LSP at the remote end. The characteristics relate to bandwidth and/or delay and/or MTU etc. The exact set of characteristics would also include available bandwidth on that TE-LSP. The end-point PE on the remote side connecting to the Tier-2 ISP's customer is also part of the update. In our case the PE-Pa and CE-B will know that to reach the 144.22.15.0/24 prefix there exists a TE-LSP from CE-A to PE-Pa. The MP-iBGP update from CE-A about the TE-LSP section from CE-A to PE-Pa also contains a label called the RSVP-stitch label. This RSVP-stitch label will have to be imposed by the upstream CE-B at the Tier-2 ISP Site-B.

2.1.1 RSVP-stitch label

When the packet to 144.22.15.1 heads from PE-Lo towards CE-B, the RSVP label for the TE-LSP to be used for that traffic is the topmost label in the packet while the VPN instance label is the second label. Assume that the PE-Lo has chosen to use the TE-LSP with the stitch option in the remote Tier-2 site, then the packets are forwarded on the TE-LSP as specified. At CE-B two things happen. The RSVP label at the head of the stack of labels is swapped with the the RSVP-stitch label.

An outer label is added which is the RSVP-splicing-LDP label propositioned by PE-Y to CE-B instead of the regular LDP label. The forwarding tables are spliced with the RSVP-splicing-LDP label to be used if the incoming labeled traffic is exiting out of the RSVP tunnel at CE-B with the RSVP-stitch label.

2.2 Across the Carrier's Core

This then carries the packet to PE-Y where the outer label which is either a LDP label or a forwarding adjacency TE-LSP RSVP label is added. This makes it four labels in the Carrier's core. Once the packet reaches PE-X the RSVP-stitch label is exposed and the packet is sent to the CE-A with the RSVP-splicing-LDP label at the top of the stack. CE-A on receiving this has already stitched the forwarding action for such a label in its forwarding table to pop the RSVP-splicing-LDP label and swap the RSVP-stitch label so that the TE-LSP section from CE-A to PE-Pa is used. The packet is then sent through the TE-LSP section thereof. This action is programmed whenever a RSVP-splicing-LDP label is the incoming label into the CE-A.

The packet then reaches the end of that TE-LSP section on to the Tier-2 Provider's Site-A customer site.

2.3 Decision at PE-Lo

The decision to send the packets for a prefix through the TE-LSP at Tier-2 Site-B is first made by PE-Lo since it has information that TE-LSP between itself and CE-B exists and that there also exists a TE-LSP section at Site-A between PE-Pa and CE-A.

2.4 Decision at CE-B

On arriving at CE-B the traffic is also subject to another decision at the CE-B as to whether to use the LDP label or the RSVP-splicing-LDP label. The latter would take the traffic through the TE-LSP section in Site-A of the Tier-2 ISP.

Thus policies can be enforced as per section 2.3 and/or 2.4. The flexibility is left to the Tier-2 ISP administrator. The choice is

two-fold.

2.5 Multiplicity of TE-LSP sections

There could be multiple TE-LSP section between CE-A and PE-Pa. These are conveyed through the MP-iBGP updates from CE-A to CE-B. For the reverse direction of traffic the vice-versa applies. So CE-B in our example could choose which one of these TE-LSP sections in Site-A could be the most appropriate. A suitable decision algorithm may be arrived at given the choices to be made at CE-B in our example.

2.6 Illustration

In other words to diagrammatically illustrate the methodology described above we provide the control and data plane exchanges for the same...

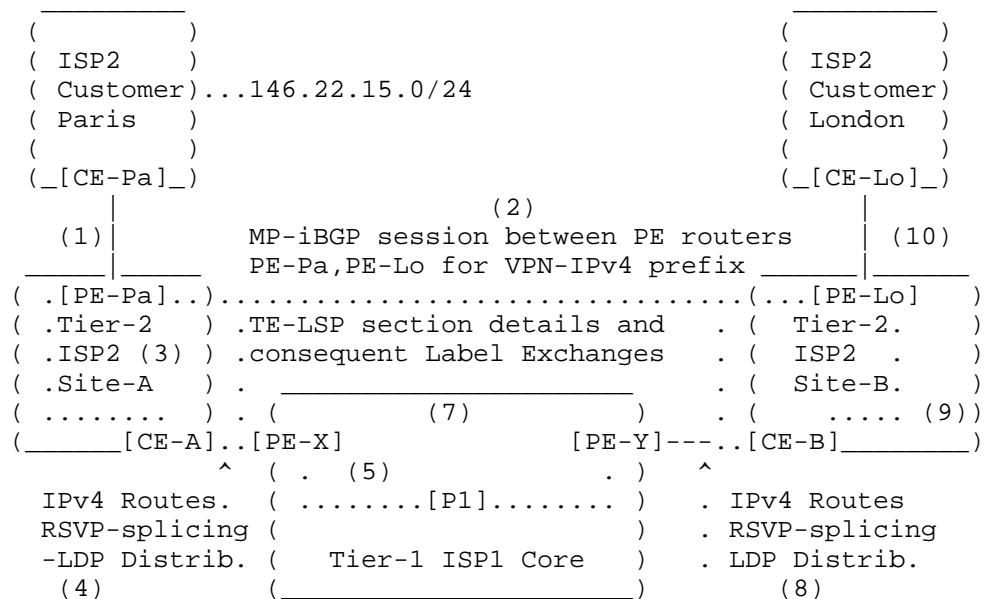


Figure 3: Reference Diagram for proposed control plane exchange for HCSc with stitch and splice TE-LSP method

Assumption is that there exist TE-LSP sections in Site-A and Site-B between CE-A and PE-Pa in the direction specified and between PE-Lo and CE-B in that specified direction. Methods to adopt Non-PHP behaviour at CE-B is to be implemented as per [RFC6511].

We demonstrate the use of the Tier-1 ISP core RSVP TE-LSP that ties together the two TE-LSP sections in Site-A and Site-B in the data plane example. This RSVP TE-LSP too has non-PHP behaviour for its egress LSR PE-X for traffic in the London to Paris direction. The same non-PHP behaviour for the RSVP TE-LSP between CE-A and PE-Pa is also in vogue.

Legend for the control plane exchange is as follows :

- (1) CE-Pa sends update for 146.22.15.0/24 to PE-Pa
- (2) An MP-iBGP update for Net=146.22.15.0/24 with next hop as PE-Pa and label assignment as 99 is sent to PE-Lo from PE-Pa
- (2.1) An MP-iBGP update for TE-LSP section between CE-A to PE-Pa describing a RSVP-stitch label 1000 with characteristics of Site-A TE-LSP is sent to CE-B and PE-Lo.
- (3) An RSVP TE-LSP between CE-A and PE-Pa with label as 500 with Non-PHP behaviour is assumed to exist
- (4) The CE-A device sends an LDP update with Net=PE-Pa with NH=CE-A and a label assignment of 12 to PE-X where the label 12 is a RSVP-splicing-LDP label. It also sends a normal LDP label for the same FEC.
- (7) An MP-iBGP update is sent from PE-X to PE-Y with Net=PE-Pa NH=PE-X and Label as 4 where label 4 is identified as a RSVP-splicing-LDP label.
- (5) An RSVP forwarding Adjacency TE-LSP is assumed to exist between PE-X and PE-Y from latter to former with non-PHP behaviour as per [RFC6511]. The labels between PE-X and P1 is 2001 and PE-Y and P1 is 2000.
- (8) An LDP update with Net=PE-Pa and NH=PE-Y with label as 11 from PE-Y to CE-B where 11 is a RSVP-splicing-LDP label. There is also an LDP update sent for normal LDP for the same FEC.
- (9) An RSVP TE-LSP between PE-Lo and CE-B with label as 400 with non-PHP behaviour is assumed to exist.
- (10) null

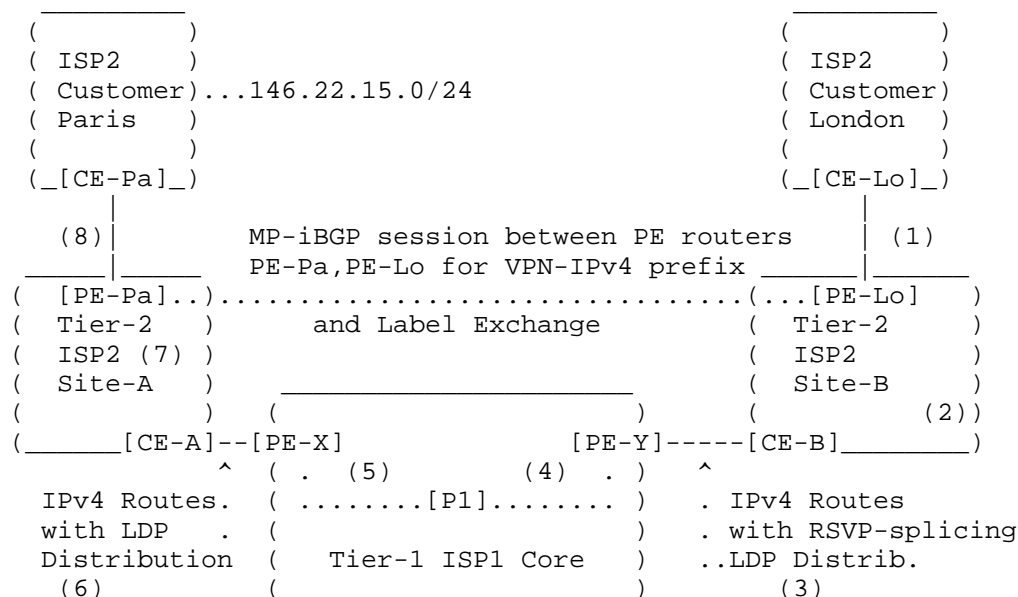


Figure 4: Reference Diagram for proposed Data Plane exchange for HCSC splicing method

Assume TE-LSPs exist between PE-Lo and CE-B in Site-B and CE-A and PE-Pa in Site-A. Also assume a forwarding adjacency LSP constructed using RSVP exists between PE-Y and PE-X in the said direction from Y to X.

- (1) IP packet with IP-DA as 146.22.15.1
- (2) Label packet with RSVP label 400,99, IP-DA 146.22.15.1
- (3) Label packet with MPLS labels 11,1000,99, IP-DA 146.22.15.1
- (4) Label packet with MPLS labels 2000,4,1000,99, IP-DA 146.22.15.1
- (5) Label packet with MPLS labels 2001,4,1000,99, IP-DA 146.22.15.1
- (6) Label packet with MPLS labels 12,1000,99, IP-DA 146.22.15.1
- (7) Label packet with MPLS labels 500,99, IP-DA 146.22.15.1
- (8) IP packet with IP-DA as 146.22.15.1

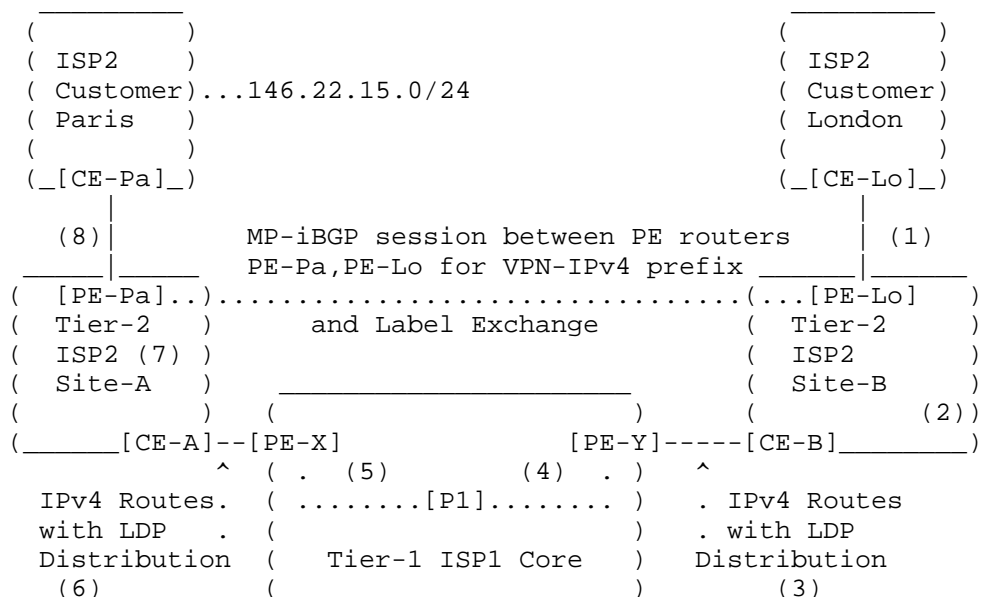


Figure 5: Reference Diagram for Data Plane exchange for proposed scheme with HCSC with LDP labels in the Carrier's Core.

Note : Control plane has not been shown for sake of brevity.

Assume TE-LSPs exist between PE-Lo and CE-B in Site-B and CE-A and PE-Pa in Site-A. Also assume there is no forwarding adjacency LSP constructed using RSVP exists between PE-Y and PE-X in the said direction from Y to X. There are only LDP labels assigned in that direction.

- (1) IP packet with IP-DA as 146.22.15.1
- (2) Label packet with RSVP label 400,99, IP-DA 146.22.15.1
- (3) Label packet with MPLS labels 11,1000,99, IP-DA 146.22.15.1
- (4) Label packet with MPLS labels 3000,4,1000,99, IP-DA 146.22.15.1
- (5) Label packet with MPLS labels 4,1000,99, IP-DA 146.22.15.1
- (6) Label packet with MPLS labels 12,1000,99, IP-DA 146.22.15.1
- (7) Label packet with MPLS labels 500,99, IP-DA 146.22.15.1
- (8) IP packet with IP-DA as 146.22.15.1

2.7 Utility

It is possible to envisage the following advantages as coming out of this proposed architecture.

- o TE-LSPs in multiple sites may be needed to be spliced together.
- o Such TE-LSPs may have been constructed to give a specific QoS for select FECs / Prefixes.
- o Without this scheme that ties the TE-LSP sections in multiple sites together, traffic may pass through TE-LSP with a given QoS in one site and may not pass through similar TE-LSP sections in other sites.
- o Splicing them together with a TE-LSP in the Tier-1 ISP gives the traffic a complete QoS experience from start to finish.
- o Multiple TE-LSP sections with different characteristics may exist in multiple sites. As per MP-iBGP updates coming in the CE devices in the Tier-2 ISP sites may choose one of them to provide the said QoS to such traffic.
- o The decision / choice to use either LDP in the sites and/or in the Carrier's core may be done by suitable algorithms that sense degradation in delay or bandwidth or a cost metric.

3 Security Considerations

No additional security considerations exist except those that apply already in the current specifications.

4 IANA Considerations

MP-iBGP PDU formats for TE-LSP characteristics and for passing the RSVP-stitch label need to be added to this document.

Changes to LDP updates to indicate plain LDP labels and RSVP-splicing-LDP labels need to be incorporated. A single bit or type / code value needs to indicate whether the LDP label exchanged is either a LDP label or a RSVP-splicing-LDP label.

These will be done in the future versions of the document.

5 References

5.1 Normative References

- [RFC4364] Rosen, E. and Y. Rekhter, "BGP/MPLS IP Virtual Private Networks (VPNs)", RFC 4364, February 2006.
- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC2205] Braden, R., Ed., Zhang, L., Berson, S., Herzog, S., and S. Jamin, "Resource ReSerVation Protocol (RSVP) -- Version 1 Functional Specification", RFC 2205, September 1997.
- [RFC3209] Awduche, D., Berger, L., Gan, D., Li, T., Srinivasan, V., and G. Swallow, "RSVP-TE: Extensions to RSVP for LSP Tunnels", RFC 3209, December 2001.
- [RFC3473] Berger, L., Ed., "Generalized Multi-Protocol Label Switching (GMPLS) Signaling Resource ReserVation Protocol-Traffic Engineering (RSVP-TE) Extensions", RFC 3473, January 2003.
- [RFC4875] Aggarwal, R., Ed., Papadimitriou, D., Ed., and S. Yasukawa, Ed., "Extensions to Resource Reservation Protocol - Traffic Engineering (RSVP-TE) for Point-to-Multipoint TE Label Switched Paths (LSPs)", RFC 4875, May 2007.

- [RFC5420] Farrel, A., Ed., 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.

5.2 Informative References

- [RFC4379] Kompella, K. and G. Swallow, "Detecting Multi-Protocol Label Switched (MPLS) Data Plane Failures", RFC 4379, February 2006.
- [RFC4761] Kompella, K., Ed., and Y. Rekhter, Ed., "Virtual Private LAN Service (VPLS) Using BGP for Auto-Discovery and Signaling", RFC 4761, January 2007.
- [RFC5920] Fang, L., Ed., "Security Framework for MPLS and GMPLS Networks", RFC 5920, July 2010.
- [RFC5921] Bocci, M., Ed., Bryant, S., Ed., Frost, D., Ed., Levrau, L., and L. Berger, "A Framework for MPLS in Transport Networks", RFC 5921, July 2010.
- [MPLSVPN-ARCH] Jim Guichard et.al, MPLS and VPN Architectures, ISBN-1-58705-002-1
- [RFC6511] Zafar Ali et.al, Non-Penultimate Hop Popping Behavior and Out-of-Band Mapping for RSVP-TE Label Switched Paths

Authors' Addresses

Bhargav Bhikkaji
Dell-Force10,
350 Holger Way,
San Jose, CA
U.S.A

Email: Bhargav_Bhikkaji@dell.com

Balaji Venkat Venkataswami
Dell-Force10,
Olympia Technology Park,
Fortius block, 7th & 8th Floor,

Plot No. 1, SIDCO Industrial Estate,
Guindy, Chennai - 600032.
TamilNadu, India.
Tel: +91 (0) 44 4220 8400
Fax: +91 (0) 44 2836 2446

EMail: BALAJI_VENKAT_VENKAT@dell.com

Shankar Raman
Department of Computer Science and Engineering
I.I.T Madras,
Chennai - 600036
TamilNadu,
India.

EMail: mjsraman@cse.iitm.ac.in

Prof.Gaurav Raina
Department of Electrical Engineering,
I.I.T Madras,
Chennai - 600036,
TamilNadu,
India.

EMail: gaurav@ee.iitm.ac.in

Internet Engineering Task Force
Internet-Draft
Intended status: Standards Track
Expires: April 23, 2013

H. Chen
Huawei Technologies
N. So
Tata Communications
A. Liu
Ericsson
L. Liu
KDDI R&D Lab Inc.
October 20, 2012

Extensions to RSVP-TE for P2MP LSP Egress Local Protection
draft-chen-mppls-p2mp-egress-protection-07.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 April 23, 2013.

Copyright Notice

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

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

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

Table of Contents

1. Introduction	3
2. Terminology	4
3. Conventions Used in This Document	4
4. Mechanism	4
4.1. An Example of Egress Local Protection	4
4.2. Set up of Backup sub LSP	5
4.3. Forwarding State for Backup sub LSP(s)	5
4.4. Detection of Egress Node Failure	6
5. Egress Local Protection with FRR	6
6. Representation of a Backup Sub LSP	7
6.1. EGRESS_BACKUP_SUB_LSP Object	7
6.1.1. EGRESS_BACKUP_SUB_LSP IPv4 Object	7
6.1.2. EGRESS_BACKUP_SUB_LSP IPv6 Object	8
6.2. EGRESS_BACKUP_SECONDARY_EXPLICIT_ROUTE Object	9
7. Path Message	9
7.1. Format of Path Message	9
7.2. Processing of Path Message	10
7.2.1. Backup LSP for One-to-One Protection	10
7.2.2. Backup LSP for Facility Protection	11
8. Processing of Resv Message	11
9. IANA Considerations	11
10. Acknowledgement	12
11. References	12
11.1. Normative References	12
11.2. Informative References	13
Authors' Addresses	13

1. Introduction

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

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

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

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

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

This document proposes a new method to locally protect the egress nodes of a P2MP LSP, which is called Egress Local Protection. It specifies the mechanism and extensions to RSVP-TE for locally protecting an egress node of a Traffic Engineered (TE) point-to-multipoint (P2MP) Label Switched Path through using a backup P2MP sub LSP. The new method overcomes the disadvantages described above. The same extensions and mechanism can also be used to protect the egress node of a TE P2P LSP.

2. Terminology

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

3. Conventions Used in This Document

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

4. Mechanism

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

4.1. An Example of Egress Local Protection

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

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

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

During normal operation, the traffic transported by the P2MP LSP is forwarded through R3 to L1, then delivered to its destination CE1. When the failure of L1 is detected, R3 forwards the traffic to the backup egress node La, which then delivers the traffic to its destination CE1. The time for switching the traffic after L1 fails

is within tens of milliseconds.

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

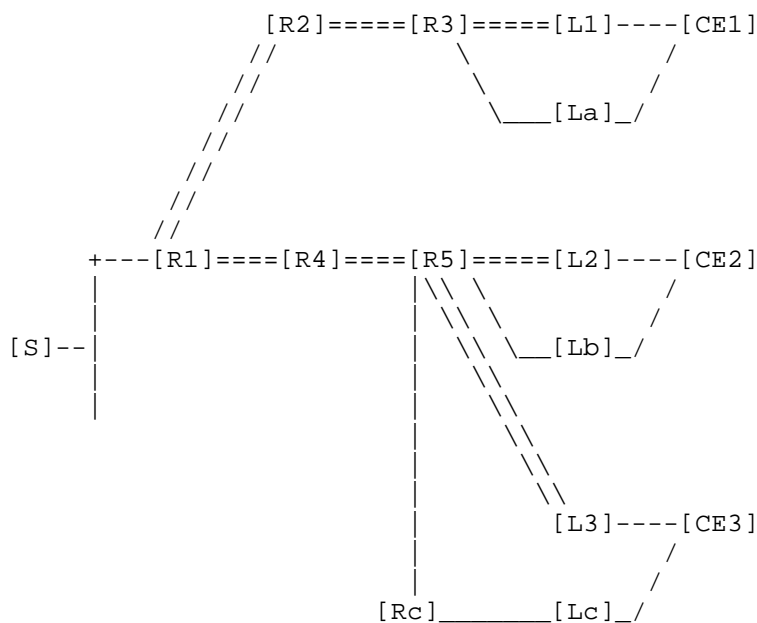


Figure 1: P2MP sub LSP for Locally Protecting Egress

4.2. Set up of Backup sub LSP

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

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

4.3. Forwarding State for Backup sub LSP(s)

The forwarding state for the backup sub LSP is different from that for a P2MP S2L sub LSP. After receiving the RSVP-TE RESV message for the backup sub LSP, the previous hop node creates a forwarding entry

with an inactive state or flag called inactive forwarding entry. This inactive forwarding entry is not used to forward any data traffic during normal operations. It SHALL only be used after the failure of the primary egress node.

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

4.4. Detection of Egress Node Failure

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

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

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

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

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

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

5. Egress Local Protection with FRR

RFC4875 "Extensions to RSVP-TE for P2MP TE LSPs" describes how to use RFC 4090 "Fast Reroute Extensions to RSVP-TE for LSP Tunnels" (FRR

for short) to locally protect failures in a link or intermediate node of a P2MP LSP. However, there is not any standard that locally protects the egresses of the P2MP LSP. The egress local protection mechanism proposed in this document fills this gap. Thus, through using the egress local protection and the FRR, we can locally protect the egress nodes, all the links and the intermediate nodes of a P2MP LSP. The traffic switchover time is within tens of milliseconds whenever any of the egresses, the links and the intermediate nodes of the P2MP LSP fails.

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

6. Representation of a Backup Sub LSP

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

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

6.1. EGRESS_BACKUP_SUB_LSP Object

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

6.1.1. EGRESS_BACKUP_SUB_LSP IPv4 Object

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

EGRESS_BACKUP_SUB_LSP Class = 50,
 EGRESS_BACKUP_SUB_LSP_IPv4 C-Type = 3

```

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

```

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

6.1.2. 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

```

      0               1               2               3
    0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
|               Egress Backup Sub LSP IPv6 destination address       |
|               (16 bytes)                                           |
|               ....                                                 |
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+
|               Egress IPv6 address                                   |
|               (16 bytes)                                           |
|               ....                                                 |
+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+

```

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.

7.2.1. Backup LSP for One-to-One Protection

If the previous hop node of the primary egress node receives the Path message with an egress backup sub LSP descriptor list and the request for protection via the one-to-one backup method, 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.

7.2.2. Backup LSP for Facility Protection

The facility backup method will be used for locally protecting a primary egress node if the previous hop node of the primary egress node receives the Path message with an egress backup sub LSP descriptor list and the request for protection via the facility backup method.

The previous hop node selects or creates a backup LSP tunnel from itself to the backup egress designated for protecting the primary egress. If there exists a backup LSP tunnel from itself to the backup egress that satisfies the constraints given in the PATH message, then this tunnel is selected; otherwise, a new backup LSP tunnel to the backup egress will be created.

After having a backup LSP tunnel, the previous hop node assigns the label allocated by the backup egress for the backup LSP as a top label (or called backup label).

When the previous hop node detects a failure in the primary egress, it has to import the traffic for the protected P2MP LSP into the backup bypass tunnel using the backup label as the top label.

8. Processing of Resv Message

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

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

9. IANA Considerations

TBD

10. Acknowledgement

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

11. References

11.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC3692] Narten, T., "Assigning Experimental and Testing Numbers Considered Useful", BCP 82, RFC 3692, January 2004.
- [RFC2205] Braden, B., Zhang, L., Berson, S., Herzog, S., and S. Jamin, "Resource ReSerVation Protocol (RSVP) -- Version 1 Functional Specification", RFC 2205, September 1997.
- [RFC3031] Rosen, E., Viswanathan, A., and R. Callon, "Multiprotocol Label Switching Architecture", RFC 3031, January 2001.
- [RFC3209] Awduche, D., Berger, L., Gan, D., Li, T., Srinivasan, V., and G. Swallow, "RSVP-TE: Extensions to RSVP for LSP Tunnels", RFC 3209, December 2001.
- [RFC3473] Berger, L., "Generalized Multi-Protocol Label Switching (GMPLS) Signaling Resource ReserVation Protocol-Traffic Engineering (RSVP-TE) Extensions", RFC 3473, January 2003.
- [RFC4090] Pan, P., Swallow, G., and A. Atlas, "Fast Reroute Extensions to RSVP-TE for LSP Tunnels", RFC 4090, May 2005.
- [RFC4875] Aggarwal, R., Papadimitriou, D., and S. Yasukawa, "Extensions to Resource Reservation Protocol - Traffic Engineering (RSVP-TE) for Point-to-Multipoint TE Label Switched Paths (LSPs)", RFC 4875, May 2007.
- [P2MP FRR] Le Roux, J., Aggarwal, R., Vasseur, J., and M. Vigoureux, "P2MP MPLS-TE Fast Reroute with P2MP Bypass Tunnels", draft-leroux-mpls-p2mp-te-bypass , March 1997.

11.2. Informative References

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

Authors' Addresses

Huaimo Chen
Huawei Technologies
Boston, MA
USA

Email: huaimo.chen@huawei.com

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

Email: ning.so@tatacommunications.com

Autumn Liu
Ericsson
CA
USA

Email: autumn.liu@ericsson.com

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

Email: le-liu@kddilabs.jp

Internet Engineering Task Force
Internet-Draft
Intended status: Standards Track
Expires: April 25, 2013

H. Chen
Huawei Technologies
N. So
Tata Communications
A. Liu
Ericsson
L. Liu
KDDI R&D Lab Inc.
October 22, 2012

Extensions to RSVP-TE for P2MP LSP Ingress Local Protection
draft-chen-mppls-p2mp-ingress-protection-07.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 April 25, 2013.

Copyright Notice

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

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

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

Table of Contents

1. Introduction	3
2. Terminology	3
3. Conventions Used in This Document	4
4. Mechanism	4
4.1. An Example of Ingress Local Protection	4
4.2. Set up of Backup P2MP sub Tree	5
4.3. Forwarding State for Backup P2MP sub Tree	5
4.4. Detection of Failure around Ingress	6
5. Ingress Local Protection with FRR	7
6. Protocol Extensions	7
6.1. New RSVP-TE Messages	8
6.1.1. LSP Information Message	8
6.1.2. Backup LSP for One-to-One Backup	9
6.1.3. Backup LSP for Facility Backup	10
6.1.4. LSP Information Confirmation Message	11
6.2. New RSVP-TE Objects	12
6.2.1. Information about Existing LSP	12
6.2.2. Desire for Locally Protecting Ingress	12
6.2.3. Backup LSP for One-to-One Backup	13
6.2.4. Backup LSP for Facility Backup	13
6.3. OSPF Opaque LSA	14
7. IANA Considerations	14
8. Acknowledgement	14
9. References	15
9.1. Normative References	15
9.2. Informative References	16
Authors' Addresses	16

1. Introduction

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

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

There exist two methods for protecting an ingress node of a P2MP LSP. The first method deploys a backup P2MP LSP from a backup ingress node to the destination nodes to protect the ingress node. The main disadvantage of this method is that the backup P2MP LSP consumes additional network bandwidth along the entire LSP paths. The impact on network efficiency can be significant in case of large P2MP deployments. In addition, the backup LSP has to be linked to the primary LSP logically at the head-end to allow the fast switching in case of ingress failure.

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

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

2. Terminology

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

3. Conventions Used in This Document

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

4. Mechanism

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

4.1. An Example of Ingress Local Protection

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

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

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

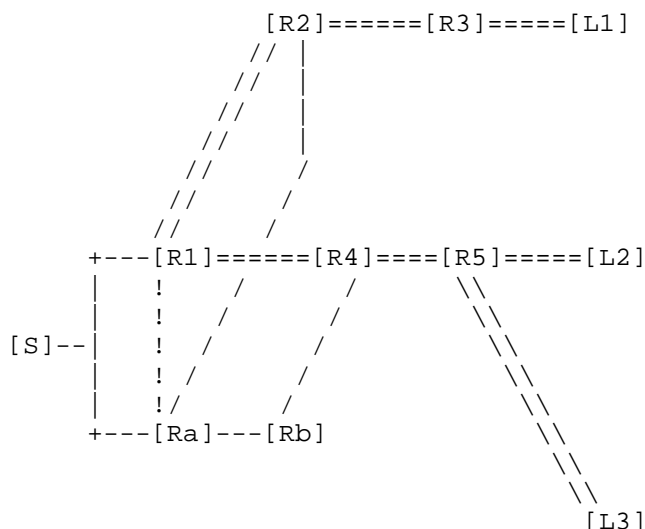


Figure 1: P2MP sub Tree for Locally Protecting Ingress

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

4.2. Set up of Backup P2MP sub Tree

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

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

4.3. Forwarding State for Backup P2MP sub Tree

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

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

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

4.4. Detection of Failure around Ingress

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

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

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

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

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

5. Ingress Local Protection with FRR

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

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

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

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

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

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

6. Protocol Extensions

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

6.1. New RSVP-TE Messages

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

LSP information messages are used to transfer the information about a P2MP LSP to a backup ingress node from an ingress node. The destination address of the LSP information message is that of the backup ingress node.

LSP information confirmation messages are used to confirm that the corresponding LSP information messages are received. In addition, the state of the backup P2MP sub tree and the action of switching over of traffic are communicated with the primary ingress through the messages.

6.1.1. LSP Information Message

6.1.1.1. Format of LSP Information Message

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

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

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

The value of the Msg Type field in the common header in the P2MP LSP

information message will be a new number to be assigned by Internet Assigned Numbers Authority (IANA).

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

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

6.1.1.2. Processing of LSP Information Message

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

When the backup ingress receives the RSVP-TE LSP information message from the primary ingress, it stores the LSP information, provides and maintains local protection for the primary ingress according to the information in the information message.

6.1.2. Backup LSP for One-to-One Backup

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

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

6.1.2.1. Construction of PATH Messages

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

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

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

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

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

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

6.1.3. Backup LSP for Facility Backup

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

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

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

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

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

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

6.1.4. LSP Information Confirmation Message

6.1.4.1. Format of LSP Information Confirmation Message

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

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

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

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

6.1.4.2. Processing of LSP Information Confirmation Message

When the backup ingress node receives a RSVP-TE LSP information message from the ingress node, it SHALL construct and send an LSP confirmation message to the ingress node to acknowledge the message received. If the backup LSP for locally protecting the primary ingress is available, the backup ingress node sets "local protection available" flag in the IPv4 (or IPv6) address sub-object of the RRO for the primary ingress and SHOULD send the updated confirmation message to the primary ingress.

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

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

6.2. New RSVP-TE Objects

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

6.2.1. Information about Existing LSP

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

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

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

6.2.2. Desire for Locally Protecting Ingress

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

primary ingress to the backup ingress.

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

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

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

6.2.3. Backup LSP for One-to-One Backup

When the backup ingress receives the PATH message with the request for Ingress Local Protection and the request for protection via the one-to-one backup method from the primary ingress, it stores the information in the message, constructs a PATH message for a backup LSP, and sends the PATH message downstream accordingly. If it has not received any PATH message from the primary ingress for an extended period of time (e.g. a cleanup timeout interval) and the BFD session between the primary ingress and backup ingress is up, it SHALL remove the information, constructs a PathTear message, and send the PathTear message downstream accordingly.

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

6.2.4. Backup LSP for Facility Backup

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

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

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

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

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

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

6.3. OSPF Opaque LSA

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

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

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

7. IANA Considerations

TBD

8. Acknowledgement

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

suggestions on this draft.

9. References

9.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC3692] Narten, T., "Assigning Experimental and Testing Numbers Considered Useful", BCP 82, RFC 3692, January 2004.
- [RFC2205] Braden, B., Zhang, L., Berson, S., Herzog, S., and S. Jamin, "Resource ReSerVation Protocol (RSVP) -- Version 1 Functional Specification", RFC 2205, September 1997.
- [RFC3031] Rosen, E., Viswanathan, A., and R. Callon, "Multiprotocol Label Switching Architecture", RFC 3031, January 2001.
- [RFC3209] Awduche, D., Berger, L., Gan, D., Li, T., Srinivasan, V., and G. Swallow, "RSVP-TE: Extensions to RSVP for LSP Tunnels", RFC 3209, December 2001.
- [RFC3473] Berger, L., "Generalized Multi-Protocol Label Switching (GMPLS) Signaling Resource ReserVation Protocol-Traffic Engineering (RSVP-TE) Extensions", RFC 3473, January 2003.
- [RFC4090] Pan, P., Swallow, G., and A. Atlas, "Fast Reroute Extensions to RSVP-TE for LSP Tunnels", RFC 4090, May 2005.
- [RFC4461] Yasukawa, S., "Signaling Requirements for Point-to-Multipoint Traffic-Engineered MPLS Label Switched Paths (LSPs)", RFC 4461, April 2006.
- [RFC4875] Aggarwal, R., Papadimitriou, D., and S. Yasukawa, "Extensions to Resource Reservation Protocol - Traffic Engineering (RSVP-TE) for Point-to-Multipoint TE Label Switched Paths (LSPs)", RFC 4875, May 2007.
- [P2MP FRR] Le Roux, J., Aggarwal, R., Vasseur, J., and M. Vigoureux, "P2MP MPLS-TE Fast Reroute with P2MP Bypass Tunnels", draft-leroux-mpls-p2mp-te-bypass , March 1997.

9.2. Informative References

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

Authors' Addresses

Huaimo Chen
Huawei Technologies
Boston, MA
USA

Email: huaimo.chen@huawei.com

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

Email: ning.so@tatacommunications.com

Autumn Liu
Ericsson
CA
USA

Email: autumn.liu@ericsson.com

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

Email: le-liu@kddilabs.jp

Network Working Group
Internet-Draft
Intended status: Informational
Expires: April 18, 2013

WQ. Cheng
L. Wang
H. Li
China Mobile
K. Liu
J. He
Huawei Technologies Co., Ltd.
F. Li
Research Institute of
Telecommunication
Transmission, China Academy of
Telecommunication Research,
MIIT. China
J. Yang
ZTE Corporation P.R.China
J. Wang
Fiberhome Telecommunication
Technologies Co., LTD
October 15, 2012

MPLS-TP Shared Ring protection (MSRP) mechanism
draft-cheng-mpls-tp-shared-ring-protection-00

Abstract

This document describes requirements and solutions for MPLS-TP Shared ring protection (MSRP) in ring topology for point to point (P2P) services. The mechanisms of MSRP are illustrated and analyzed how to satisfy the MPLS-TP requirements in RFC5654 for optimized ring protection. MSRP could support wrapping and steering protection mechanisms for P2P services, which provide a simple and reliable protection switching. The survivability of the network could be improved and the operation and maintain could be more easy. Ring protection solution for p2mp services will be documented in another draft latterly.

Status of this Memo

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

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

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

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

This Internet-Draft will expire on April 18, 2013.

Copyright Notice

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

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

Table of Contents

| | |
|--|----|
| 1. Introduction | 4 |
| 1.1. Problem statement | 4 |
| 1.1.1. Recovery for Multiple failures | 4 |
| 1.1.2. Upgrade from linear protection to ring protection | 5 |
| 1.1.3. Configuration complexity | 5 |
| 1.2. Terminology and Notation | 5 |
| 1.3. Contributing Authors | 6 |
| 2. Shared ring protection for P2P | 6 |
| 2.1. Basic conception | 6 |
| 2.1.1. The establishment of the Ring tunnels | 7 |
| 2.1.2. The distribution and management of ring labels | 8 |
| 2.1.3. Failure detection | 9 |
| 2.2. P2P wrapping | 10 |
| 2.2.1. Wrapping Link Failure | 10 |
| 2.2.2. Wrapping node Failure | 11 |
| 2.3. P2P steering | 11 |
| 3. Coordination protocol | 14 |
| 4. Conclusions and Recommendations | 14 |
| 5. IANA Considerations | 14 |
| 6. Security Considerations | 14 |
| 7. Normative References | 15 |
| Authors' Addresses | 15 |

1. Introduction

As described in 2.5.6.1. Ring Protection of MPLS-TP requirements [RFC5654], several service providers have expressed a high level of interest in operating MPLS-TP in ring topologies and require a high level of survivability function in these topologies. MPLS-TP networks are often constructed with ring topologies which allow service providers setting up a efficient and optimized ring protection mechanism to achieve simplified operation and fast recovery performance.

The requirements for MPLS-TP [RFC5654] state that recovery mechanisms are optimized for Ring topologies may be developed if it can provide following optimization scenarios:

- a. Minimize the number of OAM entities for protection
- b. Minimize the number of elements of recovery
- c. Minimize the number of labels required
- d. Minimize the amount of control and management-plane transactions
- e. Minimize the impact on information exchange if control plane supported

This document specifies MPLS-TP Shared-Ring Protection mechanisms which can meet all those requirements on Ring protection listed in [RFC5654].

This document focus on the solutions for Point-to-point transport path and a related topic in [RFC5654] states the required support for point-to-multipoint transport path. The solution for point-to-multipoint solution is under evaluation and will be illustrated in a separate document based on the basic conception in this document.

1.1. Problem statement

1.1.1. Recovery for Multiple failures

MPLS-TP is expected to be used in carrier grade metro networks and backbone networks for providing mobile backhauling, business customers' services and etc., in such kind of application scenarios the network survivability are very important.

According to R106 B in RFC5654, MPLS-TP recovery mechanisms in a ring SHOULD protect against multiple failures. It's expected deploying ring protection in ring topology could enhance the network

survivability to against multiple failures in many cases. Following context provide some more detail illustration to "multiple failures".

In metro and backbone networks, the single risk factor often affects multiple links or nodes. Some examples of risk factors are given in follows:

- multiple links using fibers in one cable or pipeline
- Several nodes shared one power supply system
- weather sensitive micro-wave system

Once one risk factor happens, multiple near-simultaneous links or nodes failures occur and those failed links or nodes may locate on single ring as well as on interconnected multiple rings. The ability of ring protecting against multiple failures should cover both multiple failures on single ring scenario and multiple failures on interconnected multiple rings.

1.1.2. Upgrade from linear protection to ring protection

It is beneficial for service providers to upgrade their MPLS-TP based network from the linear protection to ring protection without service interruption, and supporting in-service insertion and removal of a node on the ring. In order to realize this requirement, the ring protection Mechanisms should be developed and optimized to comply with this upgrading principle.

1.1.3. Configuration complexity

In the application scenarios of deploying linear protection in MPLS-TP network, the configuration of protection has close relationship with the services, LSP quantities. Especially in some large metro networks with more than ten thousands of services access node, the LSP linear protection capabilities of the metro core nodes should be large enough to meet the network planning requirements, which also leads to the complexity of network protection configurations and operations. While the ring protection is based on the mechanisms on section layer, it has loose relationship with the services quantities which could simplify the network protection configurations and operations.

1.2. Terminology and Notation

The following syntax will be used to describe the contents of the label stack:

1. The label stack will be enclosed in square brackets ("[]").
2. Each level in the stack will be separated by the '|' character.

It should be noted that the label stack may contain additional levels however, we only present the levels that are germane to the protection mechanism.

3. If the Label is assigned by Node x, the Node Name will enclosed in bracket(" ()")

1.3. Contributing Authors

Wen Ye(China Mobile)

2. Shared ring protection for P2P

None

2.1. Basic conception

This document introduces an independent logic layer of ring for both working path and protection path for shared ring protection in MPLS-TP networks. The logic layer is a ring tunnel on top of the working path or the protection path as shown in Figure 1, namely working ring tunnel or protection ring tunnel respectively. Once the ring tunnel is established, the configuration, management and protection of the ring are all based on the ring tunnel. One port can carry more than one ring tunnel, while one ring tunnel can carry several LSPs.

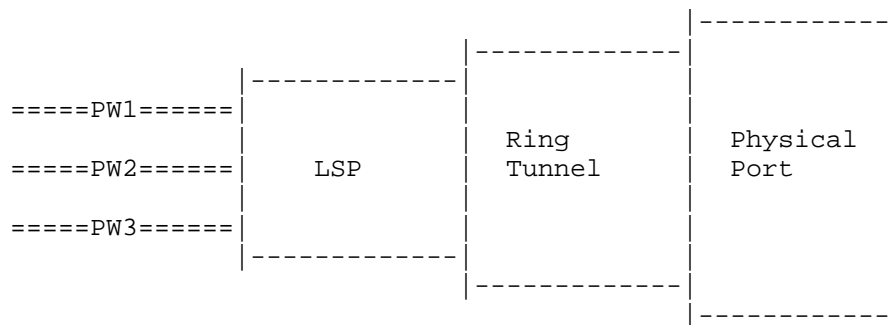


Figure 1 the logic layers of the ring

The label stack used in MPLS-TP Shared Ring Protection mechanism are shown as below.

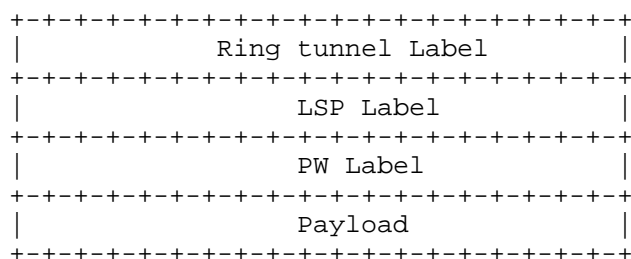


Figure 2 Label Stack used in MPLS-TP Shared Ring Protection

2.1.1. The establishment of the Ring tunnels

A ring tunnel is created as per exit node. The exit node means the node on the ring where the traffic leaves the ring. All the LSPs which transverse the ring and exit at the same ring node share the same working ring tunnels and protection ring tunnels. For each exit node, 4 ring tunnels are established:

- 1 clockwise working ring tunnel, protected by
- 1 anticlockwise protection ring tunnel,
- 1 anticlockwise working ring tunnel, protected by
- 1 clockwise protection ring tunnel.

An example is shown in Figure 3 where Node D behaves as an exit node. LSP 1 entering the ring at Node E, LSP2 at Node A and LSP 3 at Node B, all leave the ring at Node D. . To protect these LSPs traversing the ring, a clockwise working ring tunnel (RcW_D), E->F->A->B->C->D and its protection ring tunnel in the reverse direction (RaP_D), D->C->B->A->F->E->D ; Ananticlockwise working ring channel (RaW_D), C->B->A->F->E->D and its clockwise protection ring tunnel (RcP_D), D->E->F->A->B->C->D are established for Node D. Figure 3 only shows RcW_D and RaP_D for readability. The similar provisioning should be repeated for every other node on the ring. The ring tunnels created for the other nodes in Figure 3 when acting as exit node are provided as follows:

To Node A: RcW_A, RaW_A, RcP_A, RaP_A;

To Node B: RcW_B, RaW_B, RcP_B, RaP_B;

To Node C: RcW_C, RaW_C, RcP_C, RaP_C;

To Node E: RcW_E, RaW_E, RcP_E, RaP_E;

To Node F: RcW_F, RaW_F, RcP_F, RaP_F;

In Node D, two working ring tunnels, RcW_D and RaW_D are terminated; and two protection ring tunnels, RcP_D and RaP_D, are transit. That means through those working ring tunnels with protection ring tunnels, LSPs which enter the ring from Node D can reach any other nodes on the ring, while Node D can also receive the traffic from any other nodes.

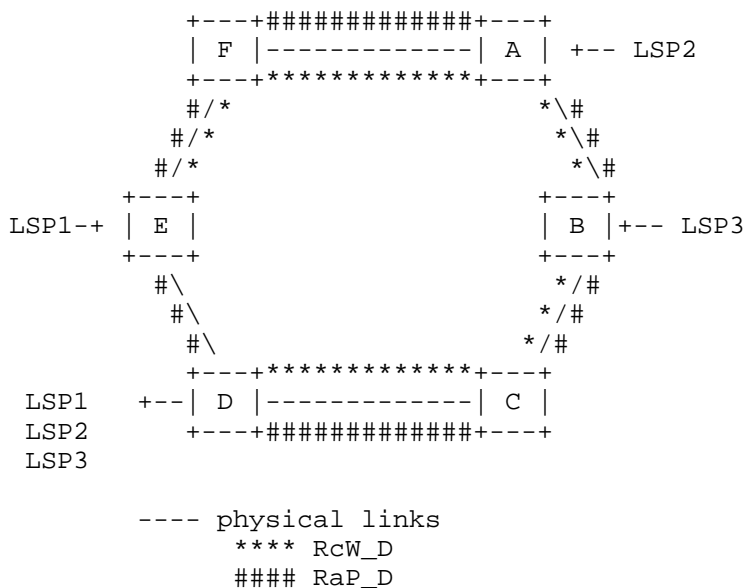


Figure 3 the Ring tunnels of the ring

2.1.2. The distribution and management of ring labels

Ring tunnel labels are distributed by means of downstream-assigned mechanism as defined in [RFC3031]. When an MPLS-TP transport path, e.g. LSP, enters the ring, according to the ring ID and the exit node, the ingress node pushes the working ring tunnel label and sends the traffic to the next hop; The transit nodes of the working ring tunnel swap the ring tunnel label and forward the packets to the next hop; When arriving at the egress node, the egress node pops the ring tunnel label and forwards the packets based on the inner LSP label and PW label. Figure 4 shows the label operation in MPLS-TP Shared Ring Protection mechanism. Suppose LSP 1 enters the ring at Node A and exit at Node D, the following label operations are executed.

1. The traffic LSP1 arrives at Node A with a label stack [LSP1] and

is supposed to be forwarded in the clockwise direction of the ring. The clockwise working ring tunnel label RcW_D will be pushed at Node A, the label stack for the forwarded packet at Node A is changed to [RcW_D(B)|LSP1]

2．Transit nodes, in this case Node B and C, forward the packet by swapping the working ring tunnel label. For Example, the label [RcW_D(B)|LSP1] is swapped to [RcW_D(C)|LSP1] at Node B.

3. When the packet arrives at Node D (i.e. egress node) with label stack [RcW_D(D)|LSP1], Node D Pops RcW_D(D) and further deals with the inner labels of LSP1.

4. All the LSPs exit at the same node share the same Ring tunnel label.

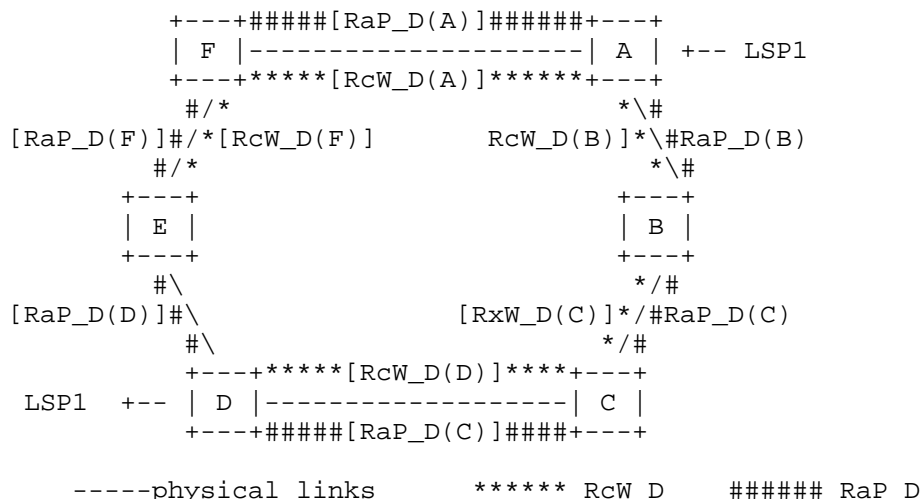


Figure 4 Label operation of the ring

2.1.3. Failure detection

The MPLS-TP section layer OAM is used to monitor the connectivity between each two adjacent nodes on the ring using the mechanisms defined in [RFC6371]. Protection switching occurs on the detection of failure on a link in the ring monitored by OAM functions.

Two end ports of a link form a MEG, and an MEG end point (MEP) function is installed in each ring port. Periodic CC-V OAM packets exchange is activated between each pair of MEPs to monitor the link

health. Sequential consecutive loss of CC-V packets (3 packets) is interpreted as a link failure.

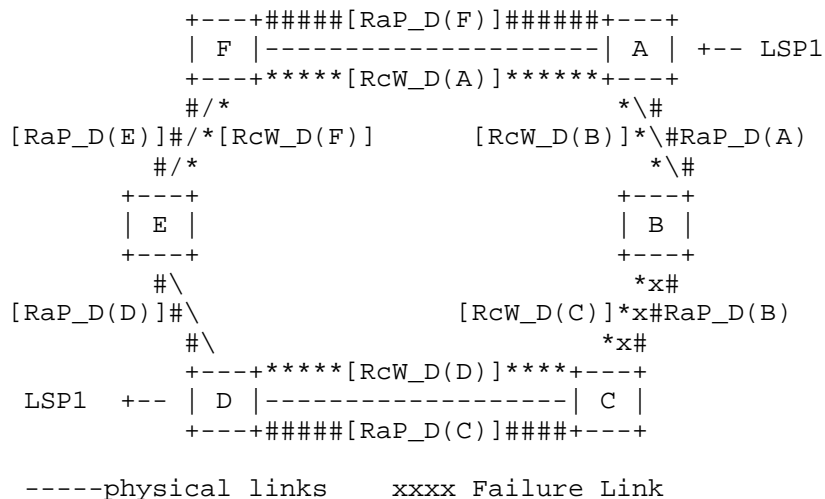
A node failure is regarded as the failure of two links attached to the node. The two nodes adjacent to the failed node detect the failure on the links connected to the failed node.

2.2. P2P wrapping

Normal state is shown in Figure 4. The clockwise LSP1 towards node D enters the ring at Node A. In normal state, LSP 1 follows the path A->B->C->D, label operation is [LSP1](original data traffic carried by LSP 1)->[RCW_D(B)|LSP1](NodeA)->[RCW_D(C)|LSP1](NodeB)->[RCW_D(D)|LSP1](NodeC)->[LSP1](data traffic carried by LSP 1). Then traffic packet will be forwarded on the basis of LSP1.

2.2.1. Wrapping Link Failure

When a link failure between Node B and Node C occurs, both Node B and Node C detect the failure by OAM mechanism. Node B switches the clockwise working ring tunnel(RcW_D) to the anticlockwise protection ring tunnel (RaP_D) and Node C switches anticlockwise protection ring tunnel(RaP_D) to the clockwise work ring tunnel(RcW_D). The data traffic which enters the ring at Node A and exits at Node D follows the path A->B->A->F->E->D->C->D. The label operation is [LSP1](Original data packet)-> [RcW_D(B)|LSP1](NodeA)-> [RaP_D(A)|LSP1](NodeB)->[RaP_D(F)|LSP1](NodeA)->[RaP_D(E)|LSP1] (NodeF)-> [RaP_D(D)|LSP1] (NodeE)-> [RaP_D(C)|LSP1] (NodeD)-> [RcW_D(D)|LSP1](NodeC)->[LSP1](Exit data packet).



***** RcW_D ##### RaP_D

Figure 5 Link Failure of P2P Wrapping

2.2.2. Wrapping node Failure

When Node B fails, Node A detects the failure between A and B and switches the clockwise work ring tunnel(RcW_D) to the anticlockwise protection ring tunnel(RaP_D), Node C detects the failure between C and B and switches the anticlockwise protection ring tunnel(RaP_D) to the clockwise working ring tunnel(RcW_D). The data traffic which enters the ring at Node D follows the path A->F->E->D->C->D. The label operation is [LSP1](original data traffic carried by LSP 1)-> [RaP_D(F)|LSP1](NodeA)->[RaP_D(E)|LSP1](NodeF)-> [RaP_D(D)|LSP1](NodeE)-> [RaP_D(C)|LSP1] (NodeD)->[RcW_D(D)|LSP1] (NodeC)->[LSP1](data traffic carried by LSP 1).

```

+---+#####[RaP_D(F)]#####+---+
| F |-----| A | +--- LSP1
+---+*****[RcW_D(A)]*****+---+
#/*                                         *\#
[RaP_D(E)]#/*[RcW_D(F)]          [RcW_D(B)]*\#RaP_D(A)
#/*                                         *\#
+---+                                     xxxxxx
| E |                                     x B x
+---+                                     xxxxxx
#\\                                     */#
[RaP_D(D)]#\\          [RcW_D(C)]*/#RaP_D(B)
#\\                                     */#
+---+*****[RcW_D(D)]*****+---+
LSP1 +--- | D |-----| C |
+---+#####[RaP_D(C)]#####+---+

-----physical links      xxxxxx  Failure Node
*****RcW_D                #####  RaP_D

```

Figure 6 Node Failure of P2P Wrapping

2.3. P2P steering

Each working ring tunnel is associated with a protection ring tunnel in the opposite direction. Every node needs to know the ring topology by configuration or topology discovery. When the failure occurs, the nodes which detect the failure will spread the fault information in the opposite direction node by node in the ring respectively. When the node receives the message that informs the

failure, it will quickly figure out the location of the fault by the topology information that is maintained by itself, so that it will determine whether the LSPs enter the ring from itself needs switch-over. If yes, it will switch the LSPs from the working ring tunnel to its protection ring tunnel. If no, it will ignore the fault indication message.

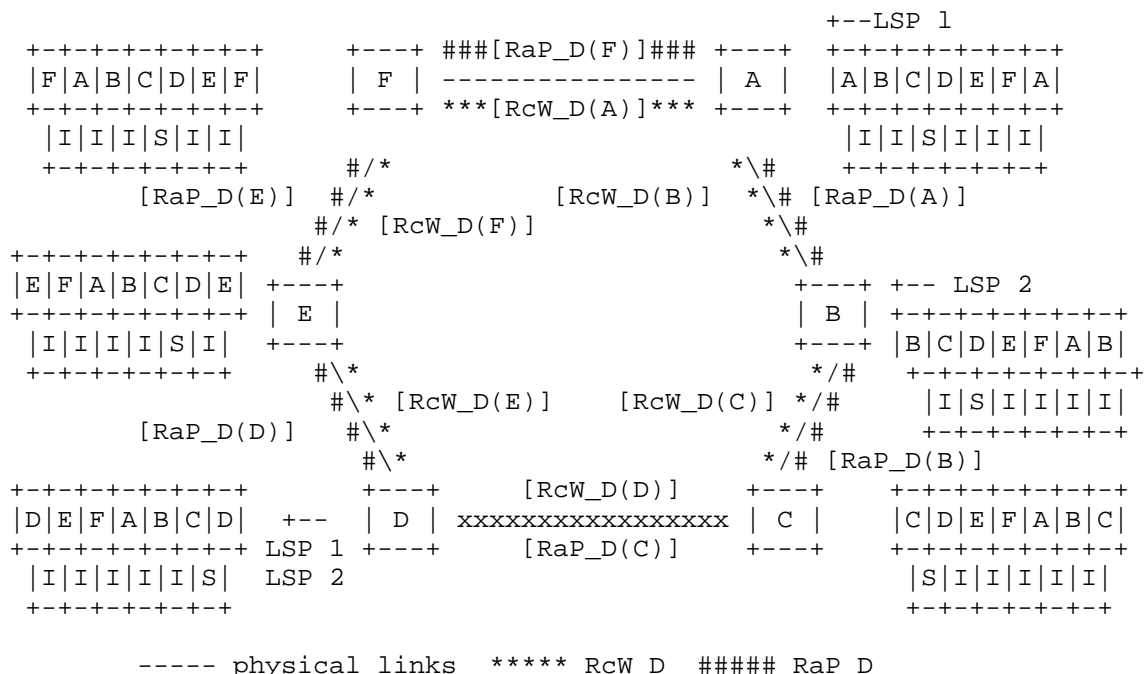


Figure 7 the P2P steering operation and protection switching(1)

Steering Example is shown in figure 7. LSP1 enters the ring from Node A while Node B has an LSP2, and both of them have the same destination node D. As per Figure 7, in normal state, LSP1 follows the path A->B->C->D, the label operation is [LSP1](original data traffic carried by LSP 1)->[RcW_D(B)]|LSP1](NodeA)->[RcW_D(C)]|LSP1](NodeB)->[RcW_D(D)]|LSP1](NodeC)->[LSP1] (data traffic carried by LSP 1) . LSP2 goes through the path B->C->D, the label operation is [LSP2]->[RcW_D(C)]|LSP2](NodeB)->[RcW_D(D)]|LSP2](NodeC)-> [LSP2] (data traffic carried by LSP 1) .

If the link between C and D breaks down, as Figure 7 shows, according to the fault detection function of each link, Node D will find out that there is a failure in the link between C and D, and it will

update the link state of its ring topology, changing the link state between C and D from normal to fault, as Figure 7 shows. In the direction that goes away from the failure point, Node D will send the state report message to Node E, informing Node E of the fault between C and D, and E will update the link state of its ring topology, changing the link state between C and D from normal to fault. And the like, the state report message is sent from node to node in the clockwise direction. Similar to Node D, Node C will spread the failure information in anti-clockwise direction.

Until Node A updates the link state of its ring topology, and knows there is a fault within its working path. And at the same time it can get the conclusion that the anticlockwise path from A to D is working all right. so that Node A will switch the LSP1 operation to the anticlockwise tunnel.

LSP1 will follow the path A->F->E->D, the label operation is
 [LSP1](original data traffic carried by LSP 1)->[RaP_D(F)|
 LSP1](NodeA)->[RaP_D(E)|LSP1](NodeF)->[RaP_D(D)|LSP1](NodeE)->[LSP1]
 (data traffic carried by LSP 1).

The same goes with LSP2 operation, when Node B updates the link state of its ring topology, and find out the working path fault, so it will stop sending the LSP2 operation in clockwise direction and switch the LSP2 to the anticlockwise protection tunnel. LSP2 goes through the path B->A->F->E->D, the label operation is [LSP2](original data traffic carried by LSP 2)-> [RaP_D(A)|LSP2](NodeB)->[RaP_D(F)|LSP2](NodeA)->[RaP_D(E)|LSP2](NodeF)->[RaP_D(D)|LSP2](NodeE)->[LSP2] (data traffic carried by LSP 2).

Imagine the ring between A and B breaks down, as figure 8 shows. Like above, Node B will find out that there is a fault in the link between A and B, and it will update the link state of its ring topology, changing the link state between A and B from normal to fault. The state report message is sent from node to node in the clockwise direction, informing every node that there is a fault between node A and B, so that every node updates the link state of its ring topology. Node A will find out the working path fault of LSP1 and switch LSP1 to protection Ring tunnel, while Node B finds out the LSP2 working path is all right and there is no need for switching.

```

+---+ LSP 1
+---+ ##### [RaP_D(F)] ##### +---+ +---+
| F | | A | | A | B | C | D | E | F | A |
+---+ *** [RcW_D(A)] **** +---+ +---+
| I | S | I | I | I | I | x | S | I | I | I | I | I |
+---+ #/ * x +---+
[ RaP_D(E) ] #/ * [ RcW_D(F) ] [ RcW_D(B) ] x [ RaP_D(A) ]
#/ * x +---+ LSP 2
+---+ +---+
| E | F | A | B | C | D | E | | E | | B | | B | C | D | E | F | A | B |
+---+ +---+
| I | I | S | I | I | I | # \ * * / # | I | I | I | I | I | S |
+---+ # \ * [ RcW_D(E) ] [ RcW_D(C) ] * / # +---+
[ RaP_D(D) ] # \ * * / # [ RaP_D(B) ]
+---+ # \ * * / # +---+
| D | E | F | A | B | C | D | +---+ *** [RcW_D(D)] *** +---+ | C | D | E | F | A | B | C |
+---+ +---+ | D | +---+ | C | +---+
| I | I | I | S | I | I | LSP1 +---+ ##### [RaP_D(C)] ##### +---+ | I | I | I | I | S | I |
+---+ LSP2 +---+ +---+
----- physical links ***** RcW_D ##### RaP_D

```

Figure 8 the P2P steering operation and protection switching(2)

3. Coordination protocol

TBD

4. Conclusions and Recommendations

TBD

5. IANA Considerations

None

6. Security Considerations

TBD

7. 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.
- [RFC5654] Niven-Jenkins, B., Brungard, D., Betts, M., Sprecher, N., and S. Ueno, "Requirements of an MPLS Transport Profile", RFC 5654, September 2009.
- [RFC6371] Busi, I. and D. Allan, "Operations, Administration, and Maintenance Framework for MPLS-Based Transport Networks", RFC 6371, September 2011.

Authors' Addresses

Weiqiang Cheng
China Mobile
No.32 Xuanwumen West Street
Beijing 100053
China

Email: chengweiqiang@chinamobile.com

Lei Wang
China Mobile
No.32 Xuanwumen West Street
Beijing 100053
China

Email: Wangleiyj@chinamobile.com

Han Li
China Mobile
No.32 Xuanwumen West Street
Beijing 100053
China

Email: Lihan@chinamobile.com

Kai Liu
Huawei Technologies Co., Ltd.
Huawei base, Bantian, Longgang District
Shenzhen 518129
China

Email: alex.liukai@huawei.com

Jia He
Huawei Technologies Co., Ltd.
Huawei base, Bantian, Longgang District
Shenzhen 518129
China

Email: hejia@huawei.com

Fang Li
Research Institute of Telecommunication Transmission, China Academy of Telecommunication Research, MIIT. China
Number 52, Huayuan street, Haidian District
Shenzhen 100191
China

Email: lifang@ritt.cn

Jian Yang
ZTE Corporation P.R.China
ZTE Industrial Zone, Liuxian Road, Xili District, Shenzhen
Shenzhen 518055
China

Email: yang.jian90@zte.com.cn

Junfang Wang
Fiberhome Telecommunication Technologies Co., LTD
No.5, Dongxin Lu, Guandong Industrial Park, Wuhan, Hubei
Wuhan 430073
China

Email: wjf@fiberhome.com.cn

MPLS
Internet-Draft
Intended status: Standards Track
Expires: April 15, 2013

D. Frost
S. Bryant
Cisco Systems
October 12, 2012

MPLS Generic Associated Channel (G-ACh) Test Session Control
draft-frost-mpls-test-session-00

Abstract

RFC 6374 defines procedures for packet loss and throughput measurement in MPLS networks. Some forms of measurement rely on the existence of a stream of test messages that flows between measurement points, from which the loss and throughput characteristics of the underlying data channel are inferred. This document presents procedures for the establishment and maintenance of such test sessions.

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 15, 2013.

Copyright Notice

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

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

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

Table of Contents

| | |
|--|----|
| 1. Introduction | 3 |
| 1.1. Terminology | 3 |
| 1.2. Requirements Language | 3 |
| 2. Overview | 3 |
| 3. Simple Session Control Protocol | 4 |
| 3.1. Message Format | 5 |
| 3.2. TLV Objects | 6 |
| 3.3. Session Setup | 9 |
| 3.4. Session Maintenance and Release | 9 |
| 4. Test Session Parameters | 10 |
| 4.1. Destination Test Identifier | 10 |
| 4.2. Source Test Identifier | 11 |
| 4.3. Packet Format | 11 |
| 4.4. Path Type | 12 |
| 4.5. Payload Size Range | 13 |
| 4.6. Maximum Transmission Rate | 13 |
| 5. Test Session Control | 13 |
| 6. Security Considerations | 14 |
| 7. IANA Considerations | 14 |
| 7.1. Allocation of Associated Channel Types | 15 |
| 7.2. Creation of MPLS Simple Session Control Protocol TLV
Registry | 15 |
| 7.3. Creation of MPLS Simple Session Control Protocol
Session Type Registry | 15 |
| 8. Normative References | 16 |
| Authors' Addresses | 16 |

1. Introduction

Procedures and protocol messages for packet loss, delay, and throughput measurement in MPLS networks are documented in [RFC6374]. Packet loss measurement, in that document, is classified as either direct or inferred: direct measurement is based on comparing transmit and receive counters for all data-plane traffic flowing over the channel, while inferred measurement is based on comparing the equivalent counters for a distinct stream of test traffic. Similarly, out-of-service throughput measurement entails validating the data-plane capacity of a channel by generating a stream of test traffic at a rate that meets or exceeds the expected capacity.

The Loss Measurement (LM) protocol defined in RFC 6374 relies on the existence of a test traffic stream when used to conduct inferred LM or out-of-service throughput measurement. This document defines procedures for the setup and control of such test streams via the MPLS Generic Associated Channel (G-ACh) [RFC5586].

1.1. Terminology

| Term | Definition |
|-------|---------------------------------|
| ----- | ----- |
| G-ACh | Generic Associated Channel |
| LM | Loss Measurement |
| SSCP | Simple Session Control Protocol |
| TLV | Type-Length-Value |

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. Overview

The objective is for a device acting as an LM querier to establish a test traffic stream to the LM responder in advance of initiating the LM session; this stream can then serve as the target of the LM operation, persisting until the operation is finished. Test session setup and maintenance proceeds according to the following process:

1. The querier determines the desired parameters for the test session, encodes them in a setup message as specified later in this document, and sends it to the responder. This message is transmitted periodically until either a response is forthcoming or a timeout occurs.

2. The responder, upon receiving the test session parameters, either accepts or rejects them. In either case, it formulates a response and sends it to the querier. The response indicates whether the session is accepted or rejected and, in the latter case, parameters that the responder considers acceptable. If the session was accepted, it is now considered "alive" at the responder, which maintains state for it until it times out or is explicitly released.
3. The querier, upon receiving the responder's message, knows whether the test session is now active. If not, it can retry the attempt using parameters the responder has indicated are acceptable. If so, it now does three things: it begins sending test traffic; it periodically sends a message refreshing/verifying the test session state; and it initiates an LM session that targets this test session.
4. The querier, when finished with the measurement operation, terminates the LM session, ceases sending test traffic, and sends an advisory message to the responder that the test session has ended.

In the remainder of this document the term "querier" is replaced by "initiator" in the context of test session control.

3. Simple Session Control Protocol

This document defines a new G-ACh protocol and associated Channel Type:

| Protocol | Channel Type |
|---------------------------------|--------------|
| Simple Session Control Protocol | 0xXXXX |

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

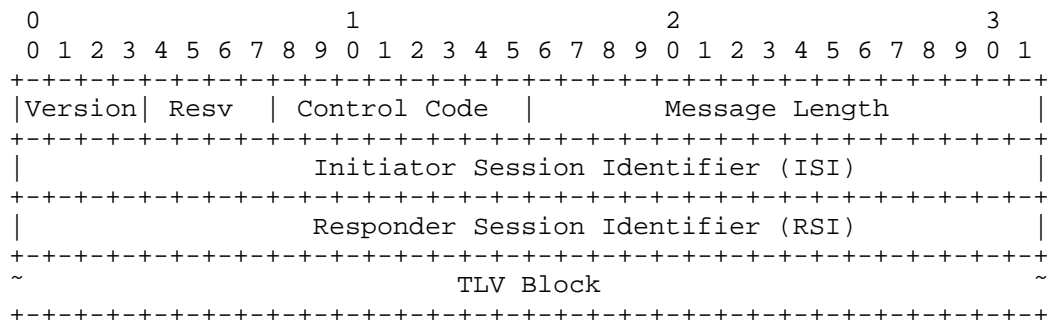
The Simple Session Control Protocol (SSCP) is a minimal "skeleton protocol" for the setup and control of point-to-point "sessions" over the G-ACh, where a session is defined abstractly as an initial agreement of application-specific parameters between the initiator and responder, followed by some form of state that is maintained between the two endpoints until either a timeout occurs or the session is explicitly released.

The only SSCP application discussed in this document is that of measurement test stream control. However, the SSCP has been defined

in a general form with the view that it may have other future applications.

3.1. Message Format

The following figure shows the format of an SSCP message, which follows the Associated Channel Header (ACH):



SSCP Message Format

Fields in this document 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.

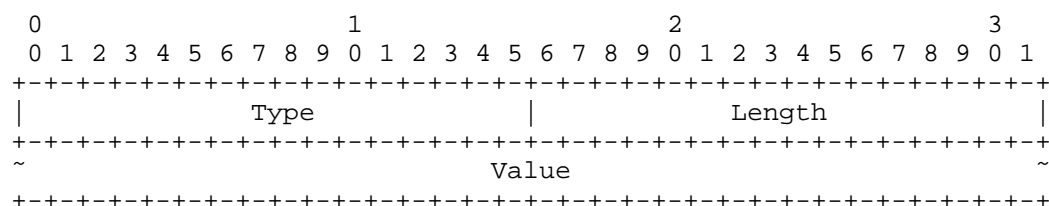
In this format, the Version field indicates the protocol version and is currently set to 0. The Message Length field indicates the size in octets of this message (i.e. of the portion of the packet that follows the Associated Channel Header). The Initiator Session Identifier (ISI) and Responder Session Identifier (RSI) fields identify the specific session to which this message belongs, via locally-significant tags allocated by the initiator and the responder respectively.

The Control Code indicates whether this message is querying, initiating, refreshing, releasing, accepting, or rejecting a session. The first four values are used by the initiator, the last two by the responder:

Code Meaning

| | |
|---|----------|
| 0 | Query |
| 1 | Initiate |
| 2 | Refresh |
| 3 | Release |
| 4 | Accept |
| 5 | Reject |

The remainder of the message consists of a sequence of Type-Length-Value (TLV) objects, which have the following format:

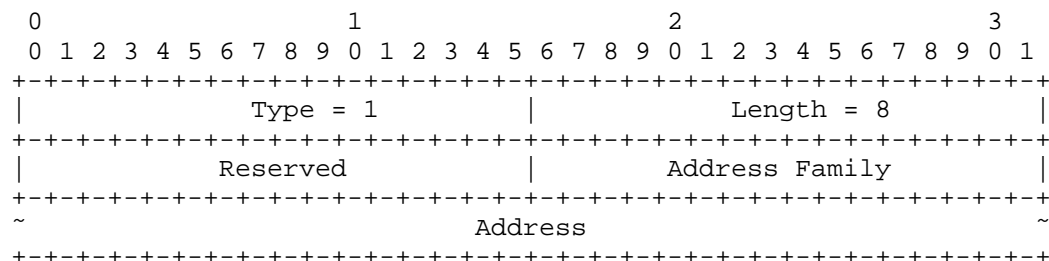


TLV Object Format

The Type field identifies the TLV Object; an IANA registry has been created to track the values of this field. Types 0-127 are reserved for use by the SSCP itself, with the rest available for application-specific allocation. The Length field specifies the length in octets of the Value field.

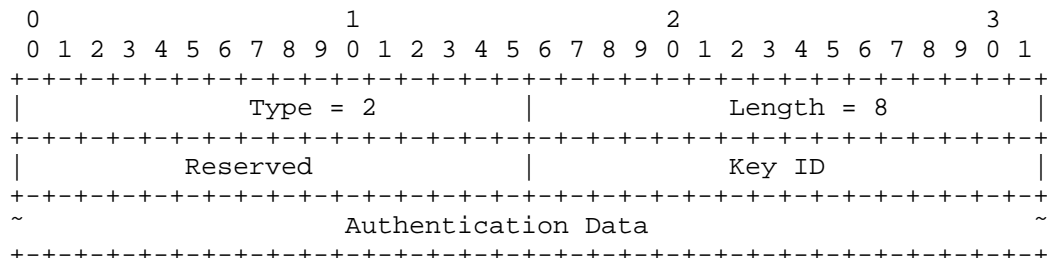
3.2. TLV Objects

3.2.1. Source Address



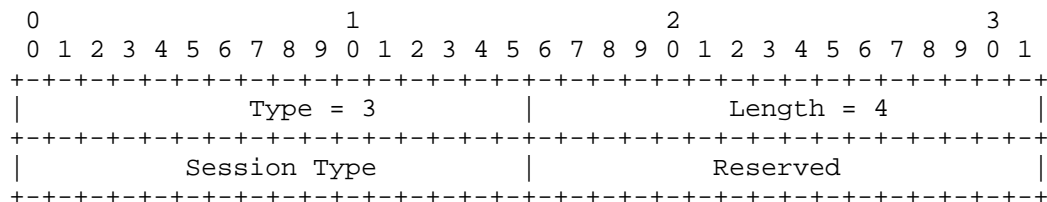
The Source Address allows the initiator to inform the responder of its address when sending an Initiate or Query message. The format of this object is identical to the Source Address TLV object described in [I-D.ietf-mpls-gach-adv].

3.2.2. Authentication



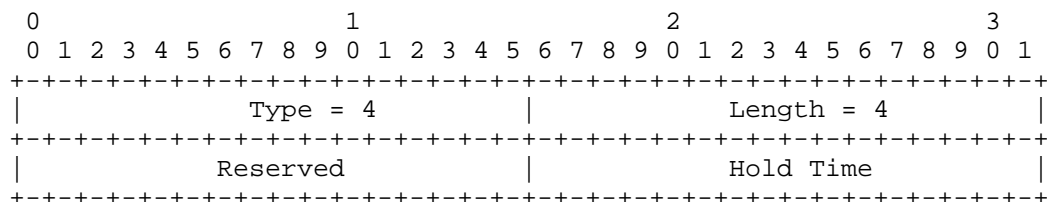
The Authentication object allows the receiver of an SSCP message to verify the identity of the message source and the integrity of the message. The format and processing semantics of this object are specified in [I-D.ietf-mpls-gach-adv].

3.2.3. Session Type



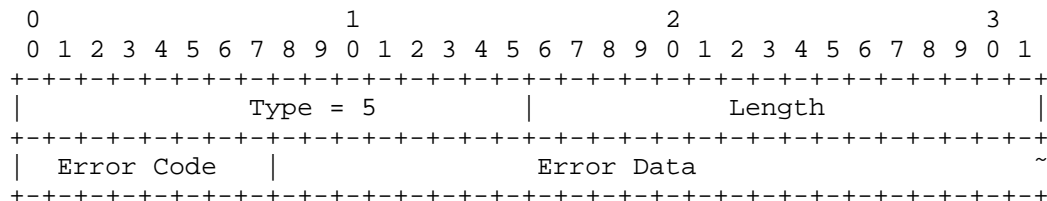
The Session Type is included in Initiate and Query messages and indicates the type of session that the initiator seeks to establish. An IANA registry has been created to track the values for the Session Type.

3.2.4. Hold Time



The Hold Time object indicates the amount of time, in seconds, the responder should keep this session alive if a refresh message is not received. When the hold timer expires, the responder discards all state associated with the session. When a refresh message is received, the responder resets its hold timer to the Hold Time.

3.2.5. Error Information



The Error Information object is included by the responder as part of a Reject message and identifies the reason for rejection in the form of an error code. Some codes may carry additional error information, in which case this information is placed in the Error Data field. The Error Data field has zero length unless otherwise noted below.

Error Code Meaning

| | |
|---|--------------------------|
| 0 | Unspecified Error |
| 1 | Protocol Error |
| 2 | Resource Unavailable |
| 3 | Unsupported Session Type |
| 4 | Unsupported Parameter |
| 5 | Authentication Failed |
| 6 | Invalid Session |

Unspecified Error: An unspecified error has prevented the requested session from being accepted. This code **MUST NOT** be used if a more specific code applies.

Protocol Error: A protocol error was found when parsing the incoming message.

Resource Unavailable: Node resources are not available to support the requested session.

Unsupported Session Type: Support for the Session Type indicated in the incoming message is not available.

Unsupported Parameter: Support for one or more of the requested session parameters is not available. The Error Data field consists of a sequence of TLV objects for the bad parameters, copied from the original request.

Authentication Failed: Authentication for the incoming message failed. A response message carrying this code **MAY** be sent as an alternative to silently dropping the offending message.

Invalid Session: The Responder Session Identifier in the incoming Refresh message is unknown or has been released.

3.3. Session Setup

The initiator begins by transmitting an Initiate message, i.e. a message with the Control Code set to Initiate. The Initiate message MUST also contain a single instance each of the Session Type and Hold Time objects.

Upon transmitting the first Initiate message, the initiator sets a retransmit timer. The message is retransmitted until either a response is received or a locally-determined timeout occurs. The retransmit period SHOULD be no shorter than three seconds.

When the responder receives an Initiate message, it determines whether it can support the requested session. If not, it sends a single Reject message to the initiator with the ISI copied from the Initiate message and with an Error Information object indicating the reason for rejection. In the case of an Unsupported Parameter error, the responder also includes a set of TLV objects that describe the parameters it supports, called the "Supported Parameters" set. This set includes the Hold Time object, which in this context indicates the longest hold time the responder supports for this session type. The other objects in the Supported Parameters set are specific to the session type.

If the responder can support the requested session, it sets the hold timer for the session to the value specified by the Hold Time object and sends a single Accept message to the initiator. The ISI of the Accept message is copied from the Initiate message, and the RSI is set at the responder's discretion.

An alternative to the above procedure is for the initiator to begin by sending a Query rather than an Initiate message. Upon receiving such a message, the responder responds with a Reject message that contains either an Error Information object or the Supported Parameters object set for this session type. The ISI of the Reject message is copied from the Initiate message.

3.4. Session Maintenance and Release

Following the acceptance of a session, the responder maintains state for the session until the session's hold timer expires or a Release message for the session is received. It MAY also terminate the session if an exceptional condition occurs; in this case it SHOULD send a Reject message to the initiator.

In order to maintain the session over time, the initiator sends periodic Refresh messages containing the RSI signaled by the responder in its most recent Accept message for this session. The responder responds to a Refresh with an Accept message containing its RSI and the ISI received in the Refresh. The refresh interval SHOULD be less than one-third of the Hold Time for the session.

When the initiator is finished with the session, it sends a Release message containing the RSI signaled by the responder in its most recent Accept message for this session. Upon receiving a Release, the responder discards all state associated with the session.

4. Test Session Parameters

This document defines the following Session Type for use in establishing test traffic streams for packet loss and throughput measurement:

| Session Type | Value |
|--------------------------|--------|
| ----- | ----- |
| Measurement Test Session | 0x0001 |

Test traffic streams are negotiated via the SSCP. This negotiation determines the format and flow characteristics of the streams.

The following subsections define the SSCP parameter objects for test sessions.

4.1. Destination Test Identifier

```

      0               1               2               3
    0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-----+-----+-----+-----+-----+-----+-----+-----+
|                                     |                                     |
|               Type = 128           |               Length = 4         |
+-----+-----+-----+-----+-----+-----+-----+-----+
|               Destination Test Identifier (DTI)                       |
+-----+-----+-----+-----+-----+-----+-----+-----+

```

The DTI is a 32-bit tag allocated by the session responder (test destination) that uniquely identifies this test stream at the destination. The session initiator (test source) includes the DTI in test packets sent to the test destination, as well as in the Target Identifier field of LM messages measuring this test stream.

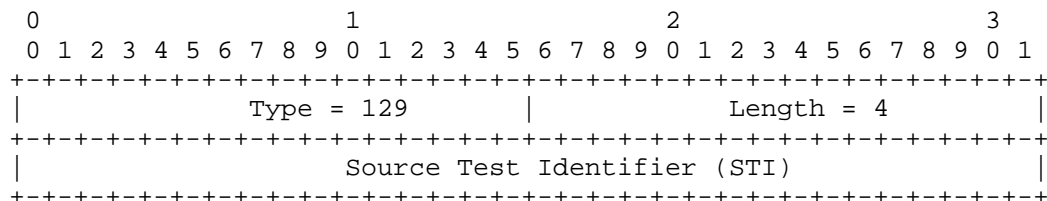
[Editor's note: An extension is proposed to the RFC 6374 LM message format whereby a block of 20 reserved bits is allocated for a "Target Identifier" field that explicitly specifies the measurement target of

this LM message, via a responder-allocated identifier such as the DTI.]

Because the Target Identifier field in the LM message format is only 20 bits long, the following restriction is placed on the DTI: If a test stream is or may be the target of an LM session, then the DTI value MUST be confined to the low-order 20 bits of the 32-bit field, and the high-order 12 bits of this field MUST be set to zero. Furthermore, the implementation MUST assume by default that this restriction is in force.

In some cases the DTI field carries an MPLS label [RFC3032]. When this is the case, the label is encoded in the low-order 20 bits of the field; the high-order 12 bits are set to zero.

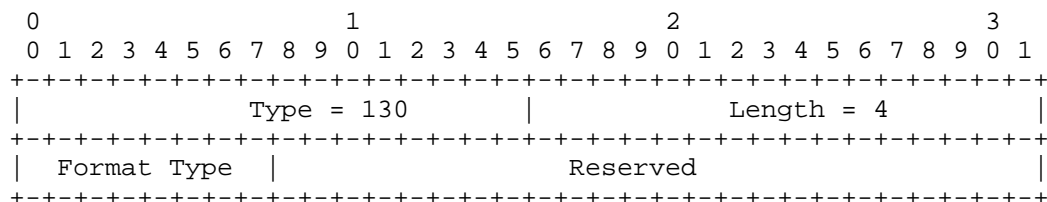
4.2. Source Test Identifier



The STI is a 32-bit tag allocated by the session initiator (test source) that uniquely identifies this test stream at the source. For bidirectional test streams, the STI replaces the DTI in the body of test messages reflected by the destination back to the source.

In some cases the STI field carries an MPLS label. When this is the case, the label is encoded in the low-order 20 bits of the field; the high-order 12 bits are set to zero.

4.3. Packet Format

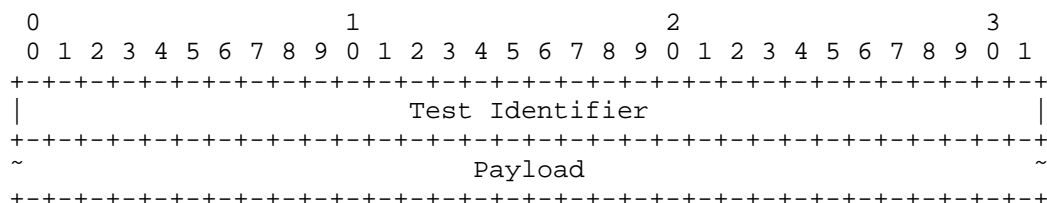


The Packet Format object identifies the format of test packets in this test stream. Possible values are:

Type Meaning

| | |
|---|------------------------------------|
| 0 | Generic Associated Channel (G-ACh) |
| 1 | MPLS Label |

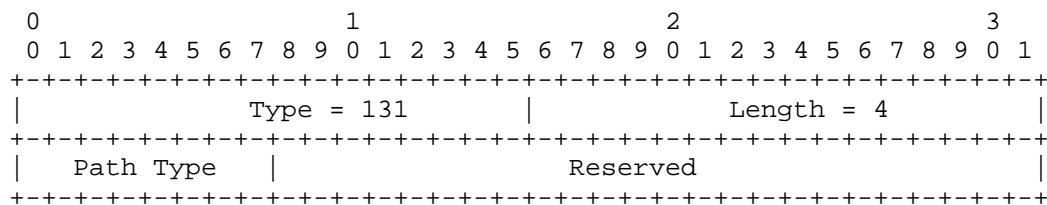
The G-ACh format indicates that test messages are sent over the G-ACh using the Channel Type allocated by IANA for test messages. In this case, test messages have the following format (after the Associated Channel Header):



In this format, the Test Identifier is set to the DTI in test messages transmitted by the test source to the test destination. In bidirectional test streams, the destination sets the Test Identifier to the STI before reflecting test messages it receives back to the source. The test message payload is set at the discretion of the test source. Support for the G-ACh format is REQUIRED.

The MPLS Label format indicates that test messages are sent as MPLS packets with a specific label at the bottom of the stack. The label values allocated by the test source and test destination are signaled via the STI and DTI objects respectively (the former only for bidirectional test streams). In this case the label serves as the test identifier; the body of the packet, i.e. the portion that follows the MPLS label stack, is considered the payload and set at the discretion of the test source.

4.4. Path Type



The Path Type object indicates whether the test stream is unidirectional or bidirectional. In a unidirectional stream, test packets are sent from the test source to the test destination and are then discarded. In a bidirectional stream, test packets are sent

from the source to the destination and reflected back to the source. Support for unidirectional sessions is REQUIRED.

Type Meaning

| | |
|---|----------------|
| 0 | Unidirectional |
| 1 | Bidirectional |

4.5. Payload Size Range

| 0 | | | | | | | | | | 1 | | | | | | | | | | 2 | | | | | | | | | | 3 | | | | | | | | | |
|------------|---|---|---|---|---|---|---|---|---|------------|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|--|--|--|--|--|--|--|--|
| 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 0 | 1 | | | | | | | | |
| Type = 132 | | | | | | | | | | Length = 4 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| Min Size | | | | | | | | | | Max Size | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |

The Payload Size Range object indicates the minimum and maximum sizes of test payloads that may be sent in this session. The sizes are specified in octets, and refer to the payload portion of test packets. For example, for the "Generic Associated Channel" test packet format the payload begins after the "Test Identifier" field, and for the "MPLS Label" format it begins after the label stack.

4.6. Maximum Transmission Rate

| 0 | | | | | | | | | | 1 | | | | | | | | | | 2 | | | | | | | | | | 3 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|--|--|--|--|--|--|--|--|
| 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 0 | 1 | | | | | | | | |
| +-----+ | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |

The Maximum Transmission Rate object indicates the maximum number of test packets per second that may be sent in this session.

5. Test Session Control

Test session control takes place according to the procedures in Section 3.3 and Section 3.4. In addition to the Session Type and Hold Time objects, the initiator MUST also include in Initiate messages a single instance each of the Payload Size Range and Maximum Transmission Rate objects. If the Packet Format object is omitted, then the "Generic Associated Channel" packet format is implied. If the Path Type object is omitted, then a unidirectional test stream is

implied. If a bidirectional test stream is requested via the Path Type, then a Source Test Identifier object MUST also be included.

If the responder can support the requested stream, it includes a Destination Test Identifier object in its Accept message.

For purposes of Unsupported Parameter errors and responses to Query messages, the Supported Parameters set includes the Packet Format, Path Type, Payload Size Range, and Maximum Transmission Rate objects.

6. Security Considerations

This document describes a simple control protocol that allows two devices to negotiate a session via the MPLS Generic Associated Channel. The most important security considerations are those that apply to securing MPLS connectivity in general; these are documented in [RFC5920]. The control protocol described in this document exchanges session data in cleartext, as this information is no more sensitive than that contained in other protocol messages that are commonly sent in cleartext. The main security considerations specific to this protocol are those concerning the verification of message authenticity and integrity, and possible denial of service.

An authentication mechanism based on cryptographic message hashing is included in the protocol, enabling receivers to verify that protocol messages were generated by a trusted source and were not corrupted or otherwise modified in transit. This mechanism also affords protection against denial-of-service attempts made by unauthorized devices. Receivers, in addition, SHOULD employ sensible rate-limiting policies to guard against the possibility of intentional or accidental denial-of-service by authorized devices. For example, implementations SHOULD anticipate the effects of receiving a large number of Initiate or Query messages within a short period of time, and take appropriate precautions to avoid resource exhaustion in such scenarios.

7. IANA Considerations

This document makes the following requests of IANA:

- o Allocation of Associated Channel Types
- o Creation of MPLS Simple Session Control Protocol TLV Registry
- o Creation of MPLS Simple Session Control Protocol Session Type Registry

7.1. Allocation of Associated Channel Types

IANA is requested to allocate an entry in the Pseudowire Associated Channel Types registry [RFC5586] for the MPLS Simple Session Control Protocol, as follows:

| Value | Description | TLV Follows | Reference |
|-------|--------------------------------------|-------------|--------------|
| (TBD) | MPLS Simple Session Control Protocol | No | (this draft) |

IANA is also requested to allocate an entry in the same registry for MPLS test messages, as follows:

| Value | Description | TLV Follows | Reference |
|-------|-------------------|-------------|--------------|
| (TBD) | MPLS Test Message | No | (this draft) |

7.2. Creation of MPLS Simple Session Control Protocol TLV Registry

IANA is requested to create a new registry, "MPLS Simple Session Control Protocol TLVs", with fields and initial allocations as follows:

| Type | Application Name | Description | Reference |
|------|---------------------------------|-----------------------------|--------------|
| 1 | Simple Session Control Protocol | Source Address | (this draft) |
| 2 | Simple Session Control Protocol | Authentication | (this draft) |
| 3 | Simple Session Control Protocol | Session Type | (this draft) |
| 4 | Simple Session Control Protocol | Hold Time | (this draft) |
| 5 | Simple Session Control Protocol | Error Information | (this draft) |
| 128 | Test Session Control | Destination Test Identifier | (this draft) |
| 129 | Test Session Control | Source Test Identifier | (this draft) |
| 130 | Test Session Control | Packet Format | (this draft) |
| 131 | Test Session Control | Path Type | (this draft) |
| 132 | Test Session Control | Payload Size Range | (this draft) |
| 133 | Test Session Control | Maximum Transmission Rate | (this draft) |

7.3. Creation of MPLS Simple Session Control Protocol Session Type Registry

IANA is requested to create a new registry, "MPLS Simple Session Control Protocol Session Types", with fields and initial allocations as follows:

| Session Type | Description | Reference |
|--------------|----------------------|--------------|
| 1 | Test Session Control | (this draft) |

8. Normative References

- [I-D.ietf-mpls-gach-adv]
Frost, D., Bryant, S., and M. Bocci, "MPLS Generic Associated Channel (G-ACh) Advertisement Protocol", draft-ietf-mpls-gach-adv-02 (work in progress), May 2012.
- [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.
- [RFC5586] Bocci, M., Vigoureux, M., and S. Bryant, "MPLS Generic Associated Channel", RFC 5586, June 2009.
- [RFC5920] Fang, L., "Security Framework for MPLS and GMPLS Networks", RFC 5920, July 2010.
- [RFC6374] Frost, D. and S. Bryant, "Packet Loss and Delay Measurement for MPLS Networks", RFC 6374, September 2011.

Authors' Addresses

Dan Frost
Cisco Systems

Email: danfrost@cisco.com

Stewart Bryant
Cisco Systems

Email: stbryant@cisco.com

Network Working Group

Internet Draft

Intended Status: Informational

Expires: April 21, 2013

X. Fu(Ed.), M. Betts, Q.

Wang

ZTE

V. Manral

Hewlett-Packard Corp.

D. McDysan (Ed.), A. Malis

Verizon

S. Giacalone

Thomson Reuters

J. Drake

Juniper Networks

October 22, 2012

Loss and Delay Traffic Engineering Framework for MPLS

draft-fuxh-mpls-delay-loss-te-framework-06

Status of this Memo

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

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

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

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

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

This Internet-Draft will expire on April 17, 2011.

Copyright Notice

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

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

Abstract

Deployment and usage of cloud based applications and services that use an underlying MPLS network are expanding and an increasing number of applications are extremely sensitive to delay and packet loss. Furthermore, in cloud computing an additional decision problem arises of simultaneously choosing the data center to host applications along with MPLS network connectivity such that the overall performance of the application is met. Mechanisms exist to measure and monitor MPLS path performance parameters for packet loss and delay, but the mechanisms work only after the path has been setup. The cloud-based and performance sensitive applications would benefit from measurement of MPLS network and potential path information that would be provided for use in the computation before LSP setup and then the selection of LSPs.

This document provides a framework and architecture to solve operator problems and requirements using current/proposed approaches, documents scalability assessment and recommendations, and identifies any needed protocol development.

Table of Contents

| | |
|---|----|
| 1. Introduction..... | 3 |
| 1.1. Scope..... | 3 |
| 2. Conventions used in this document..... | 3 |
| 2.1. Acronyms..... | 3 |
| 3. Overview of Functional Requirements..... | 4 |
| 4. Augment LSP Requestor Signaling with Performance Parameter Values
..... | 4 |
| 5. Specify Criteria for Node and Link Performance Parameter Estimation,
Measurement Methods..... | 5 |
| 6. Support Node Level Performance Information when Needed..... | 5 |
| 7. Augment Routing Information with Performance Parameter Estimates | 5 |
| 8. Augment Signaling Information with Concatenated Estimates..... | 6 |
| 9. Define Significant Performance Parameter Change Thresholds and
Frequency..... | 6 |
| 10. Define Thresholds and Timers for Links with Unusable Performance
..... | 7 |
| 11. Communicate Significant Performance Changes between Layers.... | 7 |
| 12. Support for Networks with Composite Links..... | 8 |
| 13. Support Performance Sensitive Restoration, Protection and Rerouting
..... | 8 |
| 14. Support Management and Operational Requirements..... | 8 |
| 15. Major Architectural and Scaling Challenges..... | 8 |
| 16. Approaches Considered but not Taken..... | 9 |
| 17. IANA Considerations..... | 9 |
| 18. Security Considerations..... | 9 |
| 19. References..... | 9 |
| 19.1. Normative References..... | 9 |
| 19.2. Informative References..... | 9 |
| 20. Acknowledgments..... | 10 |

1. Introduction

This draft is one of two created from draft-fuxh-mpls-delay-loss-te-framework-05 in response to comments from an MPLS Review Team (RT). This draft focuses on a framework in response to the problem statement and requirements described in a peer document [DELAY-LOSS-PS].

The purpose of this draft is to summarize a framework and architecture to meet requirements using current/proposed approaches, documents scalability assessment and recommendations, and identifies any needed protocol development.

However, computing an LSP path to meet the Network Performance Objective(NPO) for delay, loss and delay variation of these QoS classes is an open problem [DELAY-LOSS-PS]. This draft describes a framework for how the MPLS TE architecture can be augmented use information on configured, measured and/or estimated delay, loss and delay variation for use in LSP path computation and selection.

1.1. Scope

A (G)MPLS network may have multiple layers of packet, TDM and/or optical network technology and an important objective is to make a prediction of end-to-end delay, loss and delay variation based upon the current state of this network with acceptable accuracy before an LSP is established.

The (G)MPLS network may cover a single IGP area/level, may be a hierarchical IGP under control of a single administrator, or may involve multiple domains under control of multiple administrators.

An MPLS architecture for Multicast with awareness of delay, loss and delay variation will be taken up in a future version of the draft.

2. Conventions used in this document

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

2.1. Acronyms

DS-TE Differentiated Services Traffic Engineering

IGP Interior Gateway Protocol

(G)MPLS (Generalized) Multi-Protocol Label Switching

LSP Label Switched Path

RSVP-TE Resource reservation Protocol - Traffic Engineering

3. Overview of Functional Requirements

[DELAY-LOSS-PS] describes the general problem to be solved and describes a number of requirements grouped in the following subject areas for performance sensitive LSP computation and placement:

- o Augment LSP Requestor Signaling with Performance Parameter Values
- o Specify Criteria for Node and Link Performance Parameter Estimation, Measurement Methods
- o Support Node Level Performance Information when Needed
- o Augment Routing Information with Performance Parameter Estimates
- o Augment Signaling Information with Concatenated Estimates
- o Define Significant Performance Parameter Change Thresholds and Frequency
- o Define Thresholds and Timers for Links with Unusable Performance
- o Communicate Significant Performance Changes between Layers
- o Support for Networks with Composite Link
- o Support Performance Sensitive Restoration, Protection and Rerouting
- o Support Management and Operational Requirements

The following sections describe aspects of a framework for each of the above requirement sets in terms of functions, protocols and operational scenarios for meeting the requirements. In some cases the descriptions reference current/proposed potentially applicable IETF approaches. Throughout the following sections, certain scalability challenges are identified and in most cases a potential resolution approach is described - these are summarized at the end of the document.

4. Augment LSP Requestor Signaling with Performance Parameter Values

As described in [DELAY-LOSS-PS] the LSP requestor must be able to make a request for one of two types 1) a minimum possible value or 2) a maximum acceptable value for each performance parameter for each LSP.

The proposed approach [EXPRESS-PATH] within a single IGP area/level, is that only the origin (or head-end) need be aware of the required performance aspects of the LSP, since the origin has performance information for all of the candidate nodes and links from a performance parameter augmented IGP [OSPF-TE-METRIC-EXT], [ISIS-TE-METRIC-EXT].

For LSPs that traverse multiple area/levels or multiple domains, what is needed in addition to [EXPRESS-PATH] is knowledge of the node and link level performance to determine a path that meets the concatenated performance estimates as described in [DELAY-LOSS-PS]. Furthermore, information available to the LSP originator (e.g., the request type

(minimum possible value, maximum acceptable parameter value) may need to be carried in the RSVP_TE signaling message.

An alternative approach could make the performance information available to a (set of) Path Computation Elements (PCE), which the LSP requestor could consult. In this case, there would likely need to be extensions made to the PCE Protocol to carry LSP performance parameter information.

5. Specify Criteria for Node and Link Performance Parameter Estimation, Measurement Methods

Procedures to measure delay and loss on a path level between measurement points have been specified in ITU-T [Y.1731], [G.709] and [RFC 6374]. Ideally, a measurement point would occur within adjacent nodes to measure the delay, loss and delay variation performance for a combination of node and link performance. However, since this method is not universally deployed (and may never be deployed in some nodes), other methods of performance parameter estimation are needed to meet the requirements of [DELAY-LOSS-PS].

Important assumptions from [DELAY-LOSS-PS] are:

- o the timeframe of the performance parameter estimate, which is specified as the order of minutes
- o delay and loss are defined as an average and delay variation is defined based upon statistical quantiles

These assumptions could allow other methods to estimate performance parameters, such as usage of models to predict values based upon other parameters, such as load, queue thresholds and/or meters. For example, one such method could be a per QoS class based measurement from the ingress of one port to the egress of another port on a node as a function of load in a field test or laboratory to create an empirical model that could be used to insert performance parameter estimates into routing or signaling.

The switching delay on a node can be measured internally, and multiple mechanisms and data structures to do this have been defined [LEE].

6. Support Node Level Performance Information when Needed

If the IGP structure of link-level advertisements is to be used, then nodal delays can be combined with link-level performance [EXPRESS-PATH]. For example, a solution provide configuration knob to add some fixed value of a portion (e.g., one half) of node delay to link delay.

Alternatively, IGPs or a PCE information base could be extended with node-level performance parameter estimates.

7. Augment Routing Information with Performance Parameter Estimates

[DSTE-PROTO] and [EXPRESS-PATH] use information regarding bandwidth from an IGP area/level for use by performance sensitive LSPs. For a single IGP area/level, the IGP could be augmented with estimates of delay, loss

and delay variation as described in [OSPF-TE-METRIC-EXT], [ISIS-TE-METRIC-EXT]. This should also apply to a Forwarding Adjacency LSP (FA-LSP) [RFC4206]. [EXPRESS-PATH] describes how to use these augmented IGP performance measures to compute explicit paths, for example, at a path computation entity.

For LSPs that cross an IGP area/level boundary and/or traverse multiple domains, some other solution is needed for LSP path computation and selection, such as augmented PCE information bases. These PCE information bases can then be used by origin or the Path Computation engine to decide paths with the desired path properties.

Routing information could use two components to represent performance, "static" and "dynamic". The dynamic component is that caused by traffic load and queuing and would be an approximate value. The static component should be fixed and independent of load (e.g., propagation delay).

8. Augment Signaling Information with Concatenated Estimates

[DELAY-LOSS-PS] cites specific sections/appendices from [ITU-T Y.1541] regarding how performance estimates are to be composed and concatenated.

For LSPs that cross an IGP area/level boundary and/or traverse multiple domains (e.g., Autonomous Systems), if detailed performance parameter information is not provided, then one approach would be to signal the requested performance parameters for the LSP in the RSVP-TE signaling message as described in [DELAY-LOSS-RSVP-TE]. If each area/level and/or domain is unaware of the composition of performance parameters from the prior area/level and/or domains, then signaling would also need to carry the concatenation of these composed performance estimates.

Signaling information could use two components to represent performance, "static" and "dynamic". The dynamic component is that caused by traffic load and queuing and would be an approximate value. The static component should be fixed and independent of load (e.g., propagation delay).

RSVP-TE signaling across multiple area/levels or domains could include recording status of previous attempts, retries and correlation with end-end LSP performance measures to improve on a trial-and-error approach.

Another approach that could meet the requirements could be a (stateful) PCE listening to each domain, communicating amongst PCEs in other domains approximating global state to reduce probing and retries to improve scalability.

9. Define Significant Performance Parameter Change Thresholds and Frequency

In the augmented IGP approach, performance value changes should be updated and flooded in the IGP only when there is significant change in the value. The LSP originator could determine the IGP update affects performance and can decide on whether to accept the changed value, or request another computation of the LSP.

Since performance characteristics of links, nodes and FA-LSPs may change dynamically the amount of information flooded in an augmented IGP approach could be excessive and cause instability. In order to control IGP messaging and avoid being unstable when the delay, delay variation and packet loss value changes, thresholds and a limit on rate of change should be configured in the IGP control plane.

10. Define Thresholds and Timers for Links with Unusable Performance

For the extended IGP or augmented PCE information base approaches, an acceptable and unacceptable target performance value could be configured for each link (and node, if supported). This should also apply to a Forwarding Adjacency LSP (FA-LSP) [RFC4206]. If a measured or dynamically estimated (e.g., based upon load) performance value increases above the unacceptable threshold, the link (node) could be removed from consideration for future LSP path computations. If it decreases below the acceptable target value, it can then be considered for future LSP path computations.

Performance-sensitive LSPs whose path traverses links (nodes) whose performance has been deemed unacceptable by this threshold should be notified. The LSP originator can then decide if it will accept the changed performance, or else request computation of a new path that meets the performance objective.

The frequency of a link (node) changing from an unacceptable to an acceptable state should be controlled by configurable parameters.

11. Communicate Significant Performance Changes between Layers

The generic requirement is for a lower layer network to communicate significant performance changes to a higher layer network.

An end-to-end LSP (e.g., in IP/MPLS or MPLS-TP network) may traverse a FA-LSP of a server layer (e.g., an OTN ring). The boundary nodes of the FA-LSP SHOULD be aware of the performance information for this FA-LSP.

If the FA-LSP is used to form a routing adjacency and/or used as a TE link in the client network, the composition of the performance values of the links and nodes that the FA-LSP trail traverses needs to be made available for path computation. This is especially important when the performance information of the FA-LSP changes (e.g., due to a maintenance action or failure in an OTN ring).

The frequency of a lower layer network indicating a significant performance change should be controlled by configurable parameters.

A separate end-end performance measurement could be done for an LSP after it has been established (e.g., RFC 6374) if it is a lower level FA-LSP used in an LSP hierarchy. The measurement of end-to-end LSP performance may be used to inform the higher layer network of a performance parameter change.

If the performance of FA-LSP changes, the client layer must at least be notified. The client layer can then decide if it will accept the

changed performance, or else request computation of a new path that meets the performance objective.

12. Support for Networks with Composite Links

In order to assign the LSP to one of component links with different performance characteristics [CL-REQ], the RSVP-TE message could carry an indication of the request type (i.e., minimum possible value or a maximum acceptable performance parameter value) for use in component link selection or creation. The composite link should be able to take these parameters into account when assigning LSP traffic to a component link.

When Composite Links [CL-REQ] are advertised into an augmented IGP, the desirable solution is to advertise performance information for all component links into the augmented IGP [CL-FW]. Otherwise, if only partial or summarized information is advertised then the originator or a PCE cannot determine whether a computed path will meet the LSP performance objective and this could lead to crank back signaling.

13. Support Performance Sensitive Restoration, Protection and Rerouting

A change in performance of links and nodes (e.g., due to a lower level restoration action) may affect the performance of one or more end-to-end LSPs. 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. If the LSP performance objective cannot be met after a re-route is attempted, an alarm should be generated to the management plane. The solution should periodically attempt restoration for as controlled by configuration parameters to prevent excessive load on the control plane.

14. Support Management and Operational Requirements

A separate end-end performance measurement should be done for an LSP after it has been established (e.g., RFC 6374, G.709 or Y.1731). An LSP originator may re-compute a re-signal a path when the measured end-to-end performance is unacceptable. The choice by the originator to re-signal could consider a history of how accurate the performance parameter estimate is delivered by the implementation. The re-computation and re-signaling rates should be controlled by configuration parameters to prevent excessive load on the control plane.

15. Major Architectural and Scaling Challenges

As described in the preceding sections, there are a several scaling and architectural challenges, with proposed resolution as described below:

- o Frequency of performance parameter value changes limited to the order of minutes by definition
- o Augmented IGP flooding performance parameter change frequency within one area/level controlled by configuration parameters

- o Augmented PCE information base performance parameter change frequency within one area/level controlled by configuration parameters
- o Re-computation and re-signaling of LSPs whose composition of performance parameter values changes to unacceptable controlled by configuration parameters
- o Declaration of links, nodes, FA-LSPs as unacceptable/acceptable controlled by configuration parameters
- o Frequency of a lower layer network indicating a significant performance change controlled by configuration parameters
- o Re-computation and re-signaling of LSPs whose measured end-end performance is unacceptable controlled by configuration parameters

16. Approaches Considered but not Taken

One approach would be for the PCE to compute paths for use by the LSP originator for signaling. Some measurement method (e.g., RFC 6374) could then be used to measure the performance of this path. If the measurement indicates that the performance is not met then another request is made to the PCE for a different path, the originator signals for the LSP to be set up and then measured again. This "trial and error" process is very inefficient and a more predictable method is required.

17. IANA Considerations

No new IANA consideration are raised by this document.

18. Security Considerations

This document raises no new security issues.

19. References

19.1. Normative References

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

19.2. Informative References

[DELAY-LOSS-PS] X.Fu, D. McDysan et al., "Delay and Loss Traffic Engineering Problem Statement for MPLS," draft-fuxh-mpls-delay-loss-te-problem-statement

[DSTE-PROTO] Le Faucheur, F., Ed., "Protocol Extensions for Support of Diffserv-aware MPLS Traffic Engineering", RFC 4124, June 2005.

[ISIS-TE-METRIC-EXT] S. Previdi, "IS-IS Traffic Engineering (TE) Metric Extensions", draft-previdi-isis-te-metric-extensions.

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

- [EXPRESS-PATH] A. Atlas et al, "Performance-based Path Selection for Explicitly Routed LSPs", draft-atlas-mpls-te-express-path.
- [Y.1731] ITU-T Recommendation Y.1731, "OAM functions and mechanisms for Ethernet based networks", Feb 2008.
- [G.709] ITU-T Recommendation G.709, "Interfaces for the Optical Transport Network (OTN)", December 2009.
- [RFC 6374] D. Frost, S. Bryant, "Packet Loss and Delay Measurement for MPLS Networks," RFC 6374, September 2011.
- [DELAY-LOSS-RSVP-TE] X. Fu, "RSVP-TE extensions for Delay and Loss Traffic Engineering", draft-fuxh-mpls-delay-loss-rsvp-te-ext.
- [ITU-T.Y.1541] ITU-T, "Network performance objectives for IP-based services", 2011, <<http://www.itu.int/rec/T-REC-Y.1541/en>>.
- [CL-REQ] C. Villamizar, "Requirements for MPLS Over a Composite Link", draft-ietf-rtgwg-cl-requirement
- [RFC4206] Kompella, K. and Y. Rekhter, "Label Switched Paths (LSP) Hierarchy with Generalized Multi-Protocol Label Switching (GMPLS) Traffic Engineering (TE)", RFC 4206, October 2005.
- [CL-FW] C. Villamizar et al, "Composite Link Framework in Multi Protocol Label Switching (MPLS)", draft-ietf-rtgwg-cl-framework
- [LEE] Myungjin Lee , Sharon Goldberg , Ramana Rao Kompella , George Varghese "Fine-Grained Latency and Loss Measurements in the Presence of Reordering,"
<http://www.cs.bu.edu/fac/goldbe/papers/sigmet2011.pdf>

20. Acknowledgments

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

The authors would like to thank the MPLS Review Team of Stewart Bryant, Daniel King and He Jia for their many helpful comments suggestions in July 2012.

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

THIS SOFTWARE IS PROVIDED BY THE COPYRIGHT HOLDERS AND CONTRIBUTORS "AS IS" AND ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE ARE DISCLAIMED. IN NO EVENT SHALL THE COPYRIGHT OWNER OR CONTRIBUTORS BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

This code was derived from IETF RFC [insert RFC number]. Please reproduce this note if possible.

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

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

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

Intended Status: Informational

Expires: April 14, 2013

X. Fu(Ed.), M. Betts, Q.

Wang

ZTE

V. Manral

Hewlett-Packard Corp.

D. McDysan(Ed.), A. Malis

Verizon

S. Giacalone

Thomson Reuters

J. Drake

Juniper Networks

October 15, 2012

Delay and Loss Traffic Engineering Problem Statement for MPLS

draft-fuxh-mpls-delay-loss-te-problem-statement-01

Status of this Memo

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

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

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

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

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

This Internet-Draft will expire on February 27, 2013.

Copyright Notice

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

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

Abstract

Deployment and usage of cloud based applications and services that use an underlying MPLS network are expanding and an increasing number of applications are extremely sensitive to delay and packet loss. Furthermore, in cloud computing an additional decision problem arises of simultaneously choosing the data center to host applications along with MPLS network connectivity such that the overall performance of the application is met. Mechanisms exist to measure and monitor MPLS path performance parameters for packet loss and delay, but the mechanisms work only after the path has been setup. The cloud-based and performance sensitive applications would benefit from measurement of MPLS network and potential path information that would be provided for use in the computation before LSP setup and then the selection of LSPs.

This document provides a statement of problems faced by these cloud based and performance sensitive applications and describes requirements to enable the efficient and accurate measurement of the MPLS network. This also allows new performance parameters to be reported and used in the computation of MPLS services in support of these cloud based and performance sensitive applications.

Table of Contents

| | |
|--|----|
| 1. Introduction..... | 3 |
| 1.1. Scope..... | 3 |
| 2. Conventions used in this document..... | 3 |
| 2.1. Acronyms..... | 3 |
| 2.2. Terminology and Assumptions..... | 4 |
| 2.2.1. Delay..... | 4 |
| 2.2.2. Packet Loss..... | 4 |
| 2.2.3. Packet Delay Variation..... | 5 |
| 3. Motivation and Background..... | 5 |
| 3.1. General Characteristics of Performance Parameters..... | 5 |
| 3.2. Use Cases for Performance Parameter Sensitive LSP Placement | 5 |
| 4. Problem Statement..... | 6 |
| 4.1. End-to-end Measurement Insufficient for Performance Sensitive LSP Path Selection..... | 6 |
| 4.2. Lower Layer MPLS Networks Unable to Communicate Significant Performance Changes..... | 7 |
| 4.3. No Method to Communicate Significant Node/Link Performance Changes..... | 7 |
| 4.4. Routing Metrics Insufficient for Performance Sensitive Path Selection..... | 7 |
| 4.5. LSP Signaling Methods Insufficient for Performance Sensitive Path Selection..... | 8 |
| 5. Functional Requirements..... | 8 |
| 5.1. Augment LSP Requestor Signaling with Performance Parameter Values..... | 8 |
| 5.2. Specify Criteria for Node and Link Performance Parameter Estimation, Measurement Methods..... | 9 |
| 5.3. Support Node Level Performance Information when Needed.... | 9 |
| 5.4. Augment Routing Information with Performance Parameter Estimates | 10 |
| 5.5. Augment Signaling Information with Concatenated Estimates | 10 |

| | |
|---|----|
| 5.6. Define Significant Performance Parameter Change Thresholds and Frequency..... | 10 |
| 5.7. Define Thresholds and Timers for Links with Unusable Performance | 11 |
| 5.8. Communicate Significant Performance Changes between Layers..... | 11 |
| 5.9. The above requirement applies to layering with different technologies (e.g., MPLS over OTN) or to different levels within the same technology (e.g., hierarchical LSPs)..... | 11 |
| 5.10. Support for Networks with Composite Link..... | 11 |
| 5.11. Restoration, Protection and Rerouting..... | 11 |
| 5.12. Management and Operational Requirements..... | 12 |
| 6. IANA Considerations..... | 12 |
| 7. Security Considerations..... | 12 |
| 8. References..... | 12 |
| 8.1. Normative References..... | 12 |
| 8.2. Informative References..... | 12 |
| 9. Acknowledgments..... | 13 |

1. Introduction

This draft is one of two created from draft-fuxh-mpls-delay-loss-te-framework-05 in response to comments from an MPLS Review Team (RT). This draft focuses on a problem statement and requirements, for delay and loss based Traffic Engineering for MPLS networks. A peer document focuses on the framework.

The intent of this document is to focus on stating the technical aspects of the application oriented problems to be solved and specific requirements targeted to solve these problems.

It describes requirements and application needs for bounded values of delay, packet loss and delay variation.

1.1. Scope

A (G)MPLS network may have multiple layers of packet, TDM and/or optical network technology and an important objective is to make a prediction of end-to-end delay, loss and delay variation based upon the current state of this network with acceptable accuracy before an LSP is established.

The (G)MPLS network may cover a single IGP area/level, may be a hierarchical IGP under control of a single administrator, or may involve multiple domains under control of multiple administrators.

2. Conventions used in this document

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

2.1. Acronyms

SLA Service Level Agreement

SLS Service Level Specification

NPO Network Performance Objective

2.2. Terminology and Assumptions

A Service Level Agreement (SLA) is a contractual agreement that service providers have with customers for services comprised of numerical values for performance measures; for example, delay, loss and delay variation. Additionally, network operators may have Service Level Specification (SLS) that is for internal use by the operator. See [ITU-T.Y.1540], [ITU-T.Y.1541], RFC 3809, Section 4.9 [RFC3809] for examples of the form of such SLA and SLS specifications.

Network Performance Objective (NPO) is defined in section 5 of [ITU-T.Y.1541] in terms of numerical values for performance measures, principally delay, loss, and delay variation. The term NPO is used in this document since the SLA and SLS measures have network operator and service specific implications. Furthermore, the NPO measures are sufficiently well defined to address other use cases and the stated problems.

Of particular interest is the composition methods defined in Y.1541 for estimating performance parameters of candidate LSP paths based upon the performance parameter estimates/measurements of individual nodes and links.

This document assumes that the evaluation interval for a performance parameter is on the order of minutes as stated in [ITU-T Y.1541, 5.3.2], which is the same of that used in some commercial networks.

2.2.1. Delay

Section 6.2.1 of [ITU-T Y.1540] defines mean IP Packet Transfer Delay (IPTD) as the arithmetic average of the one-way delay observed between measurement points IPTD is referred to as "delay" in this document.

Section 8.2.1 of [ITU-T Y.1541] defines composition of the IPTD UNI-UNI performance parameter as "the mean IP packet transfer delay (IPTD) performance parameter, the UNI-UNI performance is the sum of the means contributed by network sections."

2.2.2. Packet Loss

Section 6.4 of [ITU-T Y.1540] defines IP Packet Loss Ratio (IPLR) as the "ratio of total lost IP packet outcomes to total transmitted IP packets in a population of interest," which is referred to as "loss" in this document.

Section 8.2.2 of [ITU-T Y.1541] defines composition of the IPLR UNI-UNI performance parameter as "may be estimated by inverting the probability of successful packet transfer across n network sections."

2.2.3. Packet Delay Variation

Section 6.2.4.2 of [ITU-T Y.1540] defines quantile-based limits on IP Packet Delay Variation (IPDV), which is referred to as "delay variation" in this document.

Section 8.2.4 of [ITU-T Y.1541] defines composition of the IPDV UNI-UNI performance parameter as "must recognize their sub-additive nature and it is difficult to estimate accurately without considerable information about the individual delay distributions." Appendix IV of [ITU-T Y.1541] gives several examples of IPDV estimate calculations.

3. Motivation and Background

3.1. General Characteristics of Performance Parameters

In general, nodes and links contribute to the performance parameters in the network. Another significant contributor is that of the host stack, but that is outside the scope of this document.

For many applications, the delay NPO is very important. In networks with wide geographic separation, propagation delay may dominate delay, while in local or metro networks nodal delay may become important.

Some link technologies (e.g., wireless, wifi, satellite) may have packet loss characteristics inherently different from those of other link technologies (e.g., fiber optic, cable) networks. Packet loss can be caused due to signal degradation/ high noise, which causes corrupted packets which in turn are discarded in the network. Furthermore, the loading of queues (congestion) may also result in packet loss.

Delay variation (sometimes also referred to as packet jitter) is important to some applications, such as interactive voice, video and/or multimedia communication, gaming, and simulations. If delay varies too much, then a playback buffer for such applications may underflow or overflow, resulting in a disruption to the application. Delay variation is caused primarily by queuing within a node. It can also be caused when packets take different paths, due to lower layer routing.

3.2. Use Cases for Performance Parameter Sensitive LSP Placement

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-delay direct market access. In certain networks, such as financial information networks (e.g. stock market data providers), network performance information (e.g. delay) is critical to data path selection. In these networks, extremely large amounts of money rest on the ability to access market data as quickly as possible and to predictably make trades faster than the competition. Using metrics such as hop count or link cost as routing may not always meet this need. In such networks it would be beneficial to be able to make path selection decisions based on performance data (such as delay) in a cost-effective and scalable way.

In other networks, for example, network-based VPNs there are in place between a customer and a provider a Service Level Agreement (SLA) which specifies performance objectives, such as delay, loss, and delay variation. In some cases these performance objectives are defined between specific customer locations. Furthermore, packets may be associated with certain classes as identified by packet header fields (e.g., IP DSCP, IEEE P-bits, MPLS TC bits) that are associated with different performance objectives. In these types of networks, the objective is to provide service that is no worse than the performance objective. A single SLA may support many customers of the same type. There is also a need to support specific SLAs, typically for very large customers who demand premium performance for which they are willing to pay a premium price.

In emerging cloud-based services, an additional decision problem where the application may be placed in a choice of more than one data center and the (G)MPLS network connectivity may also be chosen [CLO, CSO]. In these types of applications, the objective is to meet the overall performance of the application deployed in one more or more data centers. The performance of the intra- data center performance component is out of scope of this draft, but this overall cloud plus networking decision problem would benefit from a prediction of the MPLS network performance as part of path establishment.

4. Problem Statement

With the use cases in the previous section as motivation, there are several technical problems that currently standardized IETF protocols do not adequately address:

- o End-to-end Measurement Insufficient for Performance Sensitive LSP Path
- o Routing Metrics Insufficient for Performance Sensitive Path Selection
- o LSP Signaling Methods Insufficient for Performance Sensitive Path Selection
- o Lower Layer MPLS Networks Unable to Communicate Significant Performance Changes
- o No Method to Communicate Significant Node/Link Performance Changes

The following sections expand on each of these technical problem areas in more detail. Although some of the problem statements are made in terms of existing/proposed protocols, there is no intention to imply that the solution requires a revision to these protocols.

4.1. End-to-end Measurement Insufficient for Performance Sensitive LSP Path Selection

Methods exist to measure established LSP performance, e.g., [RFC 6374] for MPLS-TP, and are most useful in verifying support for an NPO. RFC 6374 specifies a mechanism to measure and monitor performance parameters for packet loss, and one-way and two-way delay, delay variation and

throughput. However, if measured performance is not met for an LSP there is not a standardized method to aid in an LSP originator or a proxy (e.g., PCE) to select a modified path that would meet the performance objective.

Therefore, there is a need to enable path computation that has access to an up to date recent performance estimate.

4.2. Lower Layer MPLS Networks Unable to Communicate Significant Performance Changes

Historically, when an IP/MPLS network was operated over a lower layer circuit switched network (e.g., SONET rings), a change in delay caused by the lower layer network (e.g., due to a maintenance action or failure) this was not known to the MPLS network. This resulted in delay affecting end user experience, sometimes violating NPO, SLS and/or SLA values and/or resulting in user complaints.

Using lower layer networks to provide restoration and grooming may be more efficient than performing packet only restoration, but the inability to communicate performance parameters, in particular delay, from the lower layer network to the higher layer network is an important problem to be solved in not only the composite link case [CL-REQ, section 4.2], but also in the case of single links connecting nodes.

In summary, Multi-layer GMPLS networks do not have a means to communicate a significant change in performance (e.g., delay) from one layer to another.

4.3. No Method to Communicate Significant Node/Link Performance Changes

Performance characteristics of links and nodes may change dynamically in response to a number of events. There is currently no way to automatically indicate which nodes and/or links have had significant performance changes to LSP originators or proxies so that they can attempt to recompute and signal a path that would meet the LSP performance objective.

4.4. Routing Metrics Insufficient for Performance Sensitive Path Selection

Optimization on a single metric does not meet the needs for all cases of performance sensitive path selection. In some cases, minimizing delay relates directly to the best customer experience (e.g., in TCP closer is faster or in financial trading the absolute minimum delay possible provides a competitive advantage). In other cases, user experience is relatively insensitive to delay, up to a specific limit at which point user perception of quality degrades significantly (e.g., interactive human voice and multimedia conferencing). A number of NPOs have a bound on point-point delay, and as long as this bound is met, the NPO is met - decreasing the delay is not necessary. In some NPOs, if the specified delay is not met, the user considers the service as unavailable. An unprotected LSP can be manually provisioned on a set of links to meet this type of NPO, but this lowers availability since an alternate route that meets the delay NPO cannot be determined.

One operational approach is to provision IP/MPLS networks over unprotected circuits and set the metric and/or TE-metric proportional to delay. This resulted in traffic being directed over the least delay path, even if this was not needed to meet an NPO or user experience objectives. This results in reduced flexibility and increased cost for network operators. However, the (TE) metric is often used to represent other information, such as link speed, economic cost or in support of ECMP (as described below) and may not be able to be set to be proportional to delay. Furthermore, if performance metrics such as loss and delay variation are to be supported in path selection, then proportional mapping is not possible.

Link attributes and LSP affinities [RFC 3209] can be used operationally to encode some information regarding performance, for example, indicating wired versus wireless, satellite versus terrestrial, etc. However, these attributes/affinities are used to encode other attributes and the 32 bit format is limiting in terms of numerical representation of performance objective parameters.

Another operational approach is to set (TE) metrics to (nearly) the same value so that LSPs are placed across multiple links using Equal Cost Multi-Path (ECMP) path selection. However, these parallel links may have markedly different performance characteristics (e.g., delay) and choice of a link that meets the performance objective is needed [CL-REQ, section 4.3].

IGP link and TE metrics are not sufficient to support performance sensitive path selection in a single IGP area/level [EXPRESS-PATH].

4.5. LSP Signaling Methods Insufficient for Performance Sensitive Path Selection

Current signaling approaches do not support inter area/level or inter-domain performance sensitive path selection. There is no standard for setting link attributes and LSP resource affinities [RFC 3209] between administrative domains, and since these have been used within some domains they are not a viable candidate to solve the aforementioned problems in this context. Augmenting an IGP with performance information does not solve the problem in these cases.

What is needed is a means for the originator/proxy of an LSP to confirm whether the estimated performance of a computed LSP path will meet the performance objective.

5. Functional Requirements

This section groups functional requirements intended to address the problems stated in the previous section into related areas.

5.1. Augment LSP Requestor Signaling with Performance Parameter Values

The solution needs to provide a means for an LSP requestor to signal performance parameter sensitive paths. The following requirements state the types of requests that are required.

The solution MUST provide a means to indicate which performance parameters are supported by the network area/level or domain.

The solution MUST provide a means for the LSP requestor to ask for the minimum possible value for each supported performance parameter.

For example, an LSP requestor may ask for an LSP that has the minimum possible value of delay.

The solution MUST provide a means for the LSP requestor to ask for a range of acceptable values for each supported performance parameter.

For example, an LSP requestor may ask for an LSP that has performance between a minimum value of delay and packet loss and a maximum value of delay and packet.

5.2. Specify Criteria for Node and Link Performance Parameter Estimation, Measurement Methods

The solution MUST provide a means to configure the one-way link and node performance parameters for delay, loss and delay variation.

The solution SHOULD provide a means to dynamically measure and/or estimate the one-way link and node performance parameters for delay, loss and delay variation.

As defined in section 2.2. , the estimation interval for the performance parameters is assumed to be on the order of minutes. The solution MUST not impact stability nor significantly increase convergence time if performance parameters change over a timeframe on the order of minutes.

5.3. Support Node Level Performance Information when Needed

There are several scenarios under which node-related performance parameters (delay, loss, delay variation) has a different level of importance:

1. The case of few nodes with large geographic separation, (e.g., trans-oceanic), where link delay alone would be a good approximation.
2. The case of many nodes with small geographic separation (e.g., interconnected nearby data centers) where node/device delay is very important but link delay may be negligible.
3. The case of some number of nodes with medium geographic separation, where usage of both link and node delay may be desirable.

The intent in case 1 is to measure the predominant delay in uncongested service provider networks, where geographic delay dominates and is on the order of milliseconds or more. The argument in cases 2 and 3 for including node-level queuing performance parameters is that it better represents the performance experienced by applications. The argument against including queuing related performance parameters is that if used in routing decisions it can result in routing instability. This tradeoff is discussed in detail in [CL-FW, Section 4.1.1].

The solution MUST define methods to include node level performance estimate information to routing protocols.

The solution MUST define methods to include node level performance estimate information to signaling protocols.

A specific deployment of the solution MAY choose to not use the node level performance estimates.

5.4. Augment Routing Information with Performance Parameter Estimates

The solution MUST provide a means to communicate performance parameters of both links and nodes as an estimate for use in performance sensitive LSP path selection within nodes of a single IGP area/level.

The solution SHOULD provide a means to communicate delay, loss and delay variation of links and nodes as a traffic engineering performance parameter for use in performance sensitive LSP path selection across a set of nodes in a hierarchy of IGP areas/levels.

5.5. Augment Signaling Information with Concatenated Estimates

The solution MUST provide a means to signal concatenated performance parameter estimates for both links and nodes as an estimate for use in performance sensitive LSP path selection traversing two or more separate administrative domains. See the terminology section for references on the concatenation method for specific performance parameters.

For example, the solution needs to support the capability to compute a route with X amount of bandwidth with less than Y ms of delay and less than Z% loss across multiple domains.

The solution MUST support the means to concatenate performance parameter estimates and report this for each traversed domain on the end-end path

The solution MUST interoperate with existing path selection and signaling methods traversing multiple domains.

5.6. Define Significant Performance Parameter Change Thresholds and Frequency

Delay, loss and delay variation measurements and/or estimates may be time varying. The solution MUST provide a means to control the advertisement rate of performance parameter estimates to avoid instability.

Any automatic LSP routing and/or load balancing solutions MUST NOT oscillate such that performance observed by users changes such that an NPO is violated. Since oscillation may cause reordering, there MUST be means to control the frequency of changing the path over which an LSP is placed.

5.7. Define Thresholds and Timers for Links with Unusable Performance

The solution MUST provide a means to configure a performance parameter threshold which defines placement of a node or link into an unusable state. The solution MUST provide a means to configure a performance parameter threshold which defines transition of a node or a link from an unusable state to a useable state. The solution MUST provide a means to control the minimum transition time between these states.

This unusable state is intended to operate on a link/node capability basis and not a global basis. Since state transition conditions are locally configured, all routers within a domain should synchronize this configuration value.

With current TE protocols, a refreshed LSP would use the most recent performance parameter estimates and may be rerouted based upon nodes or links being placed in an unusable performance state. Section 5.11. defines requirements for a desirable function where performance sensitive LSP re-routing would occur.

5.8. Communicate Significant Performance Changes between Layers

In order to support network NPOs and provide acceptable user experience, the solution MUST specify a protocol means to allow a lower layer server network to communicate performance parameters (e.g., delay, loss, delay variation) to the higher layer client network.

5.9. The above requirement applies to layering with different technologies (e.g., MPLS over OTN) or to different levels within the same technology (e.g., hierarchical LSPs).

5.10. Support for Networks with Composite Link

An LSP may traverse a network with Composite Links [CL-REQ]. The solution's selection of performance sensitive paths SHOULD be compatible with the general availability, stability and transient response requirements of [CL-REQ, Section 4.1].

When an LSP traverses a network with composite links that has component links provided by lower layer networks, the solution MUST interoperate with the requirements [CL-REQ, Section 4.2].

When an LSP traverses a network with composite links that has parallel component links with different characteristics, the solution MUST interoperate with the requirements [CL-REQ, Section 4.3].

5.11. Restoration, Protection and Rerouting

The ability to re-route an LSP if one or more NPO objectives are not met is highly desirable. The solution SHOULD support the capability to configure an LSP as capable of implementing performance sensitive re-routing, as detailed in the following conditional requirements.

If performance sensitive re-routing is implemented, the solution MUST provide a means to configure performance parameter threshold crossing and time values.

If performance sensitive re-routing is implemented, the solution MUST support a configuration option to move an end-to-end LSP away from any link or node whose performance violates the configured threshold.

If implemented, the solution MUST provide a means to control the frequency of LSP rerouting to avoid instability.

If performance sensitive re-routing is implemented, and revertive behavior to a preferred LSP is supported, then the preferred LSP MUST not be released. When the end-to-end performance of the preferred LSP becomes acceptable, the service is restored to this preferred LSP.

The delay performance of pre-defined protection or dynamic reroutable LSP MUST be defined by the solution in terms of the maximum acceptable delay difference between the primary and protection/restoration path MUST be specifiable in the solution. For example, [MPLS-TP-USE-CASE] defines a Relative Delay Time which is the difference of the Absolute Delay between the primary and protection path.

5.12. Management and Operational Requirements

Existing management and diagnostic protocols MUST be able to operate over networks supporting performance sensitive LSP placement.

If performance sensitive re-routing is implemented, and end-to-end measurements of the LSP performance are made, then the LSP requestor is able to request path placement for a performance sensitive LSP using the previously stated requirements. Since a threshold crossing of the end-to-end performance measurement may or may not correspond to a change in the concatenated performance parameter estimates, making any automatic decision on this basis MUST not create instability.

6. IANA Considerations

No new IANA consideration are raised by this document.

7. Security Considerations

This document raises no new security issues.

8. References

8.1. Normative References

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

8.2. Informative References

[CL-UC] C. Villamizer et al, "Composite Link Use Cases and Design Considerations," draft-ietf-rtgwg-cl-use-cases-01

- [CL-REQ] C. Villamizar et al, "Requirements for MPLS Over a Composite Link", draft-ietf-rtgwg-cl-requirement-08 .
- [CL-FW] C. Villamizar et al, "Composite Link Framework in Multi Protocol Label Switching (MPLS)", work in progress
- [ITU-T.Y.1540] ITU-T, "Internet protocol data communication service - IP packet transfer and availability performance parameters", 2011, <<http://www.itu.int/rec/T-REC-Y.1540/en>>.
- [ITU-T.Y.1541] ITU-T, "Network performance objectives for IP-based services", 2011, <<http://www.itu.int/rec/T-REC-Y.1541/en>>.
- [RFC3809] Nagarajan, A., "Generic Requirements for Provider Provisioned Virtual Private Networks (PPVPN)", RFC 3809, June 2004.
- [CLO] Young Lee et al, "Problem Statement for Cross-Layer Optimization," work in progress.
- [CSO] Greg Bernstein, Young Lee, "Cross Stratum Optimization Use-cases," work in progress.
- [EXPRESS-PATH] A. Atlas, "Performance-based Path Selection for Explicitly Routed LSPs", work in progress.
- [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.
- [MPLS-TP-USE-CASE] L. Fang, "MPLS-TP Applicability; Use Cases and Design", draft-ietf-mpls-tp-use-cases-and-design-01 .

9. Acknowledgments

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

The authors would like to thank the MPLS Review Team of Stewart Bryant, Daniel King and He Jia for their many helpful comments suggestions in July 2012.

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

THIS SOFTWARE IS PROVIDED BY THE COPYRIGHT HOLDERS AND CONTRIBUTORS "AS IS" AND ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE ARE DISCLAIMED. IN NO EVENT SHALL THE COPYRIGHT OWNER OR CONTRIBUTORS BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

This code was derived from IETF RFC [insert RFC number]. Please reproduce this note if possible.

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

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

Malcolm Betts
ZTE
Email: malcolm.betts@zte.com.cn

Qilei Wang
ZTE
Email: wang.qilei@zte.com.cn

John Drake
Juniper Networks
Email: jdrake@juniper.net

MPLS Working Group
Internet Draft

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

Intended status: Informational

Expires: January 8, 2013

July 9, 2012

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

Status of this Memo

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

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

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

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

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

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

Copyright Notice

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

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

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

Abstract

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

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

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

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

Table of Contents

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

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

1. Introduction

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

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

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

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

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

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

2. Conventions used in this document

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

2.1. Terminology

LSP Label Switched Path

2.2. Definitions

None

3. P2MP OAM

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

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

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

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

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

3.1. OAM functions for proactive monitoring

3.1.1. Continuity Check and Connectivity Verification

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

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

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

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

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

3.1.2. Remote Defect Indication

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

3.1.3. Alarm Reporting

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

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

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

3.1.4. Lock Reporting

FFS

3.1.5. Packet Loss Measurement

FFS

3.1.6. Packet Delay Measurement

FFS

3.1.7. Client Failure Indication

FFS

3.2. OAM functions for on-demand monitoring

3.2.1. Connectivity verification

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

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

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

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

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

3.2.2. Packet loss measurement

FFS

3.2.3. Diagnostic tests

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

Other requirements are ffs.

3.2.4. Route Tracing

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

Other requirements are ffs.

3.2.5. Packet delay measurement

FFS

3.3. OAM functions for administration control

3.3.1. Lock Instruct

FFS.

4. Security Considerations

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

5. IANA Considerations

There are no IANA actions required by this draft.

6. References

6.1. Normative References

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

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

6.2. Informative References

None

7. Acknowledgments

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

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

Authors' Addresses

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

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

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

Internet Engineering Task Force
Internet-Draft
Intended status: Standards Track
Updates: RFC4379
Expires: April 25, 2013

Q. Zhao
Huawei Technology
L. Fang
C. Zhou
Cisco Systems
L. Li
China Mobile
N. So
Tata Communications
K. Raza
Cisco Systems
October 22, 2012

LDP Extensions for Multi Topology Routing
draft-ietf-mpls-ldp-multi-topology-05.txt

Abstract

Multi-Topology (MT) routing is supported in IP networks with the use of MT aware IGP protocols. In order to provide MT routing within Multiprotocol Label Switching (MPLS) Label Distribution Protocol (LDP) networks new extensions are required. This document updates RFC4379.

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

Status of this Memo

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

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

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

This Internet-Draft will expire on April 6, 2013.

Copyright Notice

Copyright (c) 2012 IETF Trust and the persons identified as the

document authors. All rights reserved.

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

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

Table of Contents

| | |
|---|----|
| 1. Terminology | 4 |
| 2. Introduction | 4 |
| 3. Signaling Extensions | 5 |
| 3.1. Topology-Scoped Forwarding Equivalence Class (FEC) | 5 |
| 3.2. New Address Families: MT IP | 5 |
| 3.3. LDP FEC Elements with MT IP AF | 6 |
| 3.4. IGP MT-ID Mapping and Translation | 7 |
| 3.5. LDP MT Capability Advertisement | 8 |
| 3.6. Procedures | 9 |
| 3.7. LDP Sessions | 10 |
| 3.8. Reserved MT ID Values | 10 |
| 4. MT Applicability on FEC-based features | 10 |
| 4.1. Typed Wildcard FEC Element | 10 |
| 4.2. End-of-LIB | 11 |
| 4.3. LSP Ping | 11 |
| 4.3.1. New FEC Sub-Types | 11 |
| 4.3.2. MT LDP IPv4 FEC Sub-TLV | 12 |
| 4.3.3. MT LDP IPv6 FEC Sub-TLV | 12 |
| 4.3.4. Operation Considerations | 13 |
| 5. Error Handling | 13 |
| 5.1. MT Error Notification for Invalid Topology ID | 13 |
| 6. Backwards Compatibility | 14 |
| 7. MPLS Forwarding in MT | 14 |
| 8. Security Consideration | 14 |
| 9. IANA Considerations | 14 |
| 10. Contributors | 16 |
| 11. Acknowledgement | 17 |
| 12. References | 17 |
| 12.1. Normative References | 17 |
| 12.2. Informative References | 18 |
| Appendix A. Appendix | 18 |
| A.1. Requirements | 18 |
| A.2. Application Scenarios | 18 |
| A.2.1. Simplified Data-plane | 18 |
| A.2.2. Using MT for P2P Protection | 19 |
| A.2.3. Using MT for mLDP Protection | 19 |
| A.2.4. Service Separation | 19 |
| A.2.5. An Alternative inter-AS VPN Solution | 20 |
| Authors' Addresses | 20 |

1. Terminology

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

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

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

2. Introduction

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

- o A CSP may want to assign varying QoS profiles to traffic, based on a specific MT.
- o Separate routing and MPLS domains may be used to isolated multicast and IPv6 islands within the backbone network.
- o Specific IP address space could be routed across an MT based on security or operational isolation requirements.
- o Low latency links could be assigned to an MT for delay sensitive traffic.
- o Management traffic could be separated from customer traffic using multiple MTs, where the management traffic MT does not use links that carries customer traffic.

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

This document also updates RFC4379 by defining two new FEC types for LSP ping.

3. Signaling Extensions

3.1. Topology-Scoped Forwarding Equivalence Class (FEC)

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

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

3.2. New Address Families: MT IP

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

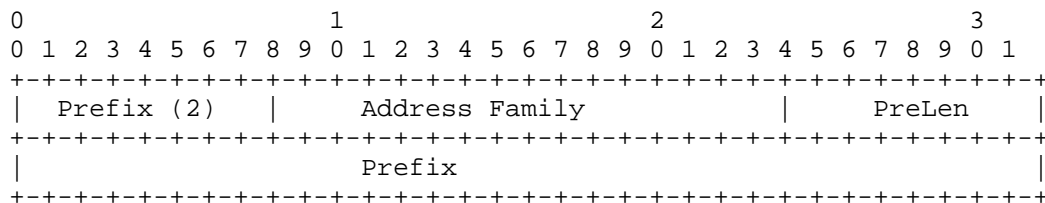


Figure 1: Prefix FEC Element Format [RFC5036]

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

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

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

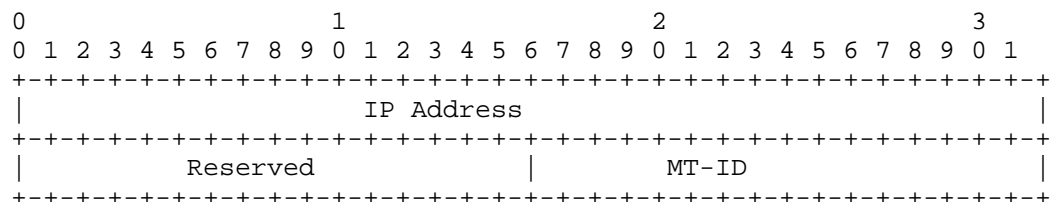


Figure 2: MT IP Address Family Format

Where "IP Address" is an IPv4 and IPv6 address/prefix for "MT IP" and "MT IPv6" AF respectively, and the field "MT-ID" corresponds to 16-bit Topology ID for given address.

Where 16-bit "MT-ID" field defines the Topology ID, and the definition and usage of the rest fields in the FEC Elements are same as defined for IP/IPv6 AF. The value of MT-ID 0 corresponds to default topology and MUST be ignored on receipt so as to not cause any conflict/confusion with existing non-MT procedures.

The proposed FEC Elements with "MT IP" Address Family can be used in any LDP message and procedures that currently specify and allow the use of FEC Elements with IP/IPv6 Address Family.

[Editors Note - RFC[5036] doesn't specify the handling of unknown Address Family. After we have introduced the two new address family here, RFC[5036] need to be updated to add the handling procedure for the unknown address families.

3.3. LDP FEC Elements with MT IP AF

The following section specifies the format extensions of the existing LDP FEC Elements. The "Address Family" of these FEC elements will be set to "MT IP" or "MT IPv6".

The MT Prefix FEC element encoding is as follows:

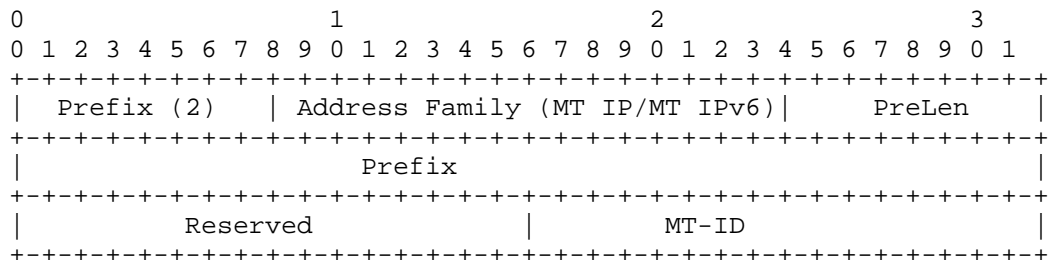


Figure 3: MT Prefix FEC Element Format

Similarly, the MT mLDP FEC elements encoding is as follows, where the mLDP FEC Type can be P2MP(6), MP2MP-up(7), and MP2MP-down(8):

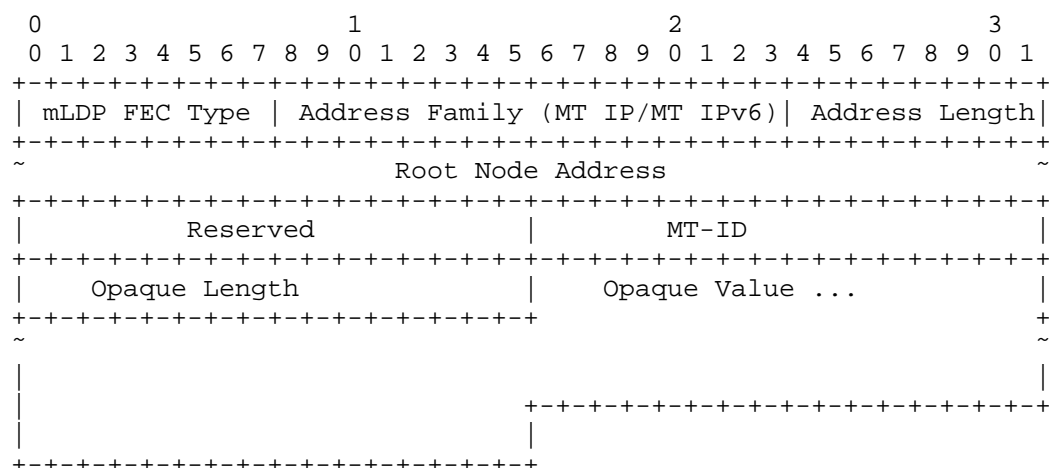


Figure 4: MT mLDP FEC Element Format

The MT Typed Wildcard FEC element encoding is as follows:

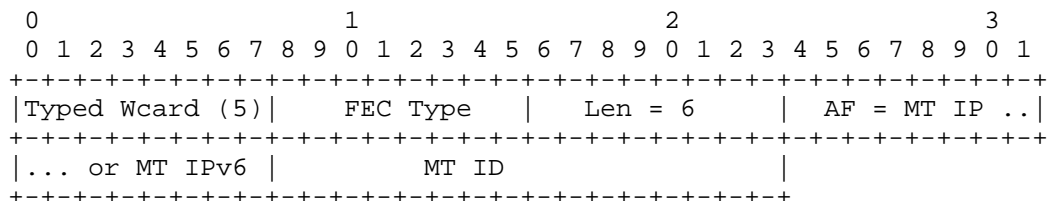


Figure 5: MT Typed Wildcard FEC Element

3.4. IGP MT-ID Mapping and Translation

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

3.5. LDP MT Capability Advertisement

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

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

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

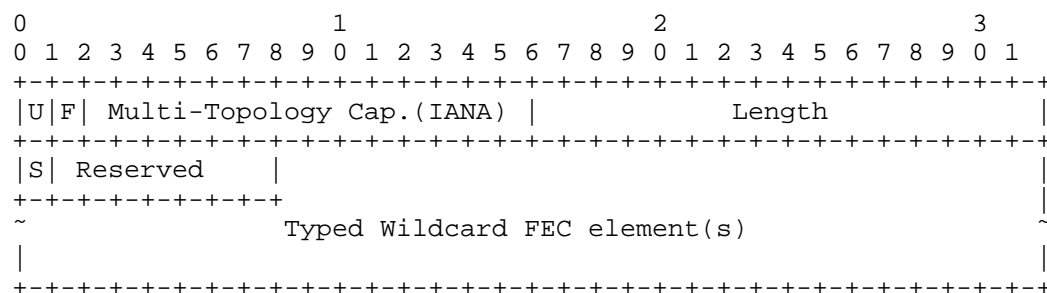


Figure 6: Multi-Topology Capability TLV Format

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

- o The encoding of Typed Wildcard FEC element, as defined in [RFC5561], is defined in the section 4.1 (Typed Wildcard FEC Element) of this document.

3.6. Procedures

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

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

3.7. LDP Sessions

If a single global label space is supported, there will be an LDP session supported for each pair of peers, regardless of the number of MTs supported between 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 recommended to establish multiple sessions for multiple topologies between these two peers. In this case, multiple LSR-IDs will need to be allocated so that each multiple topology can have its own label space ID.

3.8. Reserved MT ID Values

Certain MT topologies are assigned to serve predetermined purposes:

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

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

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

In Section 9. (IANA Considerations) this document proposes a new IANA registry "LDP Multi-Topology ID Name Space" under IANA "LDP Parameter" namespace to keep an LDP MT-ID reserved value.

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

4. MT Applicability on FEC-based features

4.1. Typed Wildcard FEC Element

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

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

The proposed format in Section 4.1 (Typed Wildcard FEC Element) allows an LSR to perform wildcard FEC operations under the scope of a topology. If an LSR wishes to perform wildcard operation that applies to all topologies, it can use a "Wildcard Topology" MT-ID. For example, upon local configuration of topology "x", an LSR may send a wildcard label withdraw request with MT-ID "x" to withdraw all its labels from the peer that advertized under the scope of topology "x". Additionally, upon a global configuration change, an LSR may send a wildcard label withdraw with the MT-ID set to "Wildcard Topology" to withdraw all its labels under all topologies from the peer.

4.2. End-of-LIB

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

4.3. LSP Ping

[RFC4379] defines procedures to detect data-plane failures in MPLS LSPs via LSP ping. The specification defines a "Target FEC Stack" TLV that describes the FEC stack being tested. This TLV is sent in an MPLS echo request message towards LSPs egress LSR, and is forwarded along the same data path as other packets belonging to the FEC.

"Target FEC Stack" TLV contains one or more sub-TLVs pertaining to different FEC types. Section 3.2 of [RFC-4379] defines Sub-Types and format for the FEC. To support LSP ping for MT LDP LSPs, this document proposes following extensions to [RFC-4379].

4.3.1. New FEC Sub-Types

We define two new FEC types for LSP ping:

- o MT LDP IPv4 FEC
- o MT LDP IPv6 FEC

We also define following new sub-types for sub-TLVs to specify these FECs in the "Target FEC Stack" TLV of [RFC-4379]:

| Sub-Type | Length | Value Field |
|----------|--------|--------------------|
| ----- | ----- | ----- |
| 24 | 5 | MT LDP IPv4 prefix |
| 25 | 17 | MT LDP IPv6 prefix |

Figure 7: new sub-types for sub-TLVs

The rules and procedures of using these sub-TLVs in an MPLS echo request message are same as defined for LDP IPv4/IPv6 FEC sub-TLV types in [RFC-4379].

4.3.2. MT LDP IPv4 FEC Sub-TLV

The format of "MT LDP IPv4 FEC" sub-TLV to be used in a "Target FEC Stack" [RFC4379] is:

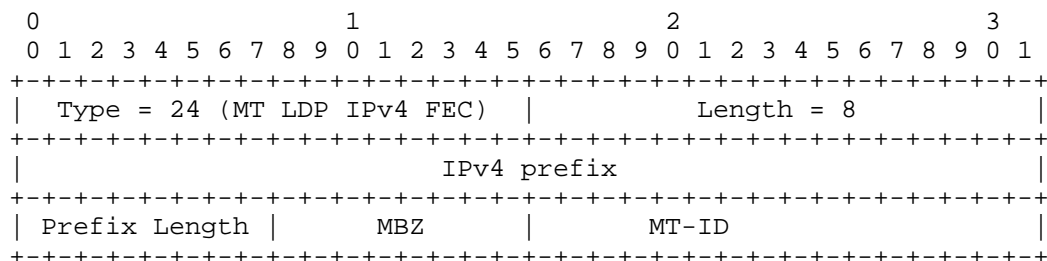


Figure 8: MT LDP IPv4 FEC sub-TLV

The format of this sub-TLV is similar to LDP IPv4 FEC sub-TLV as defined in [RFC-4379]. In addition to "IPv4 prefix" and "Prefix Length" fields, this new sub-TLV also specifies MT-ID (Multi-Topology ID). The Length for this sub-TLV is 5.

4.3.3. MT LDP IPv6 FEC Sub-TLV

The format of "MT LDP IPv6 FEC" sub-TLV to be used in a "Target FEC Stack" [RFC4379] is:

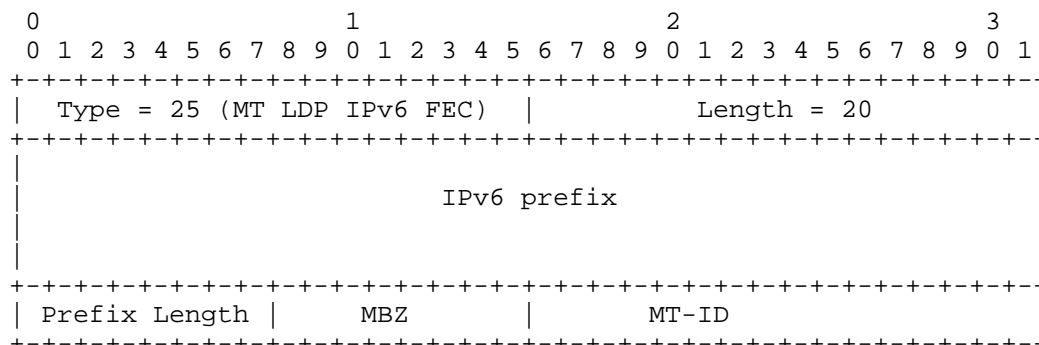


Figure 9: MT LDP IPv6 FEC sub-TLV

The format of this sub-TLV is similar to LDP IPv6 FEC sub-TLV as defined in [RFC-4379]. In addition to "IPv6 prefix" and "Prefix Length" fields, this new sub-TLV also specifies MT-ID (Multi-Topology ID). The Length for this sub-TLV is 17.

4.3.4. Operation Considerations

When detect data-plane failures using LSP Ping for a specific topology, the router will initiate an LSP Ping request with the target FEC stack containing LDP MT IP Prefix Sub-TLV in the Echo Request packet. The Echo Request packet is sent with the label binded to the IP Prefix in the topology. Once the echo request packet reaches the target router, it will process the packet and perform checks for the LDP MT IP Prefix sub-TLV present in the Target FEC Stack as described in [RFC4379] and respond according to [RFC4379] processing rules. For the case that the LSP ping with return path not specified, the reply packet may go through the default topology instead of the topology where the Echo Request goes through.

5. Error Handling

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

5.1. MT Error Notification for Invalid Topology ID

If an LSR has advertised an MT Capability TLV using the Initialization message or Capability message, which includes Typed Wildcard FEC elements with specific MT-IDs, and it receives an MT

message with a MT-ID which is not included in the supported list, it should response this "Invalid Topology ID" status code.

6. Backwards Compatibility

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

7. MPLS Forwarding in MT

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

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

8. Security Consideration

No specific security issues with the proposed solutions are known. The proposed extensions in this document do not introduce any new security considerations beyond that already apply to the base LDP specification [RFC5036] and [RFC5920].

9. IANA Considerations

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

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

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

Figure 10: New Status Codes for LDP Multi Topology Extensions

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

Figure 11: Address Family Numbers

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

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

Figure 12: LDP Multi-Topology (MT) ID Name Space

- o New Sub-TLV Types for LSP ping: Following new sub-type values under TLV type 1 (Target FEC Stack) from "Multi-Protocol Label Switching (MPLS) Label Switched Paths (LSPs) Ping Parameters" registry, and "TLVs and sub-TLVs" sub-registry.

| Sub-Type | Value Field |
|----------|--------------------|
| ----- | ----- |
| 24 | MT LDP IPv4 prefix |
| 25 | MT LDP IPv6 prefix |

Figure 13: New Sub-TLV Types for LSP ping

10. Contributors

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

Email: rtorvi@juniper.net

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

Email: huaimochen@huawei.com

Emily Chen
2717 Seville Blvd, Apt 1205,
Clearwater, FL 33764
US

Email: emily.chen220@gmail.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

Daniel King
Old Dog Consulting

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

Zhenbin Li
Huawei Technology
2330 Central Expressway
Santa Clara, CA 95050
US

Email: zhenbin.li@huawei.com

11. Acknowledgement

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

12. References

12.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC3692] Narten, T., "Assigning Experimental and Testing Numbers Considered Useful", BCP 82, RFC 3692, January 2004.
- [RFC5036] Andersson, L., Minei, I., and B. Thomas, "LDP Specification", RFC 5036, October 2007.
- [RFC5919] Asati, R., Mohapatra, P., Chen, E., and B. Thomas, "Signaling LDP Label Advertisement Completion", RFC 5919, August 2010.
- [RFC5918] Asati, R., Minei, I., and B. Thomas, "Label Distribution Protocol (LDP) 'Typed Wildcard' Forward Equivalence Class (FEC)", RFC 5918, August 2010.
- [RFC6388] Wijnands, IJ., Minei, I., Kompella, K., and B. Thomas, "Label Distribution Protocol Extensions for Point-to-Multipoint and Multipoint-to-Multipoint Label Switched Paths", RFC 6388, November 2011.
- [RFC4379] Kompella, K. and G. Swallow, "Detecting Multi-Protocol Label Switched (MPLS) Data Plane Failures", RFC 4379, February 2006.

12.2. Informative References

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

[IANA-LSPV]
Multi-Protocol Label Switching (MPLS) Label Switched Paths (LSPs) Ping Parameters, "<http://www.iana.org/assignments/mppls-lsp-ping-parameters/mppls-lsp-ping-parameters.xml>".

Appendix A. Appendix

A.1. Requirements

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

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

A.2. Application Scenarios

A.2.1. Simplified Data-plane

IGP-MT requires additional data-plane resources maintain multiple forwarding for each configured MT. On the other hand, MPLS-MT does not change the data-plane system architecture, if an IGP-MT is mapped to an MPLS-MT. In case MPLS-MT, incoming label value itself can

determine an MT, and hence it requires a single NHLFE space. MPLS-MT requires only MT-RIBs in the control-plane, no need to have MT-FIBs. Forwarding IP packets over a particular MT requires either configuration or some external means at every node, to map an attribute of incoming IP packet header to IGP-MT, which is additional overhead for network management. Whereas, MPLS-MT mapping is required only at the ingress-PE of an MPLS-MT LSP, because of each node identifies MPLS-MT LSP switching based on incoming label, hence no additional configuration is required at every node.

A.2.2. 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 distributes 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.

A.2.3. Using MT for mLDP Protection

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

A.2.4. Service Separation

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

A.2.5. 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.

Authors' Addresses

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

Email: quintin.zhao@huawei.com

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

Email: lufang@cisco.com

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

Email: czhou@cisco.com

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

Email: lilianyuan@chinamobile.com

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

Email: ning.so@tatacommunications.com

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

Email: E-mail: skraza@cisco.com

Network Working Group
Internet-Draft
Updates: 3209, 3473 (if approved)
Intended status: Standards Track
Expires: April 25, 2013

K. Kompella
Contrail Systems
October 22, 2012

Multi-path Label Switched Paths Signaled Using RSVP-TE
draft-kompella-mppls-rsvp-ecmp-02.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 April 25, 2013.

Copyright Notice

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

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

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

Table of Contents

| | |
|--|----|
| 1. Introduction | 3 |
| 1.1. Terminology | 3 |
| 1.2. Conventions used in this document | 4 |
| 2. Theory of Operation | 5 |
| 2.1. Multi-path Label Switched Paths | 5 |
| 2.2. ECMP | 6 |
| 2.3. Discussion | 8 |
| 2.4. The Capabilities of TE-based Load Balancing | 9 |
| 3. Operation of MLSPs | 10 |
| 3.1. Signaling MLSPs | 10 |
| 3.1.1. MLSP_TUNNEL Sender Template | 10 |
| 3.1.2. MLSP_TUNNEL Filter Specification | 11 |
| 3.2. Label Allocation | 11 |
| 3.3. Bandwidth Accounting | 11 |
| 3.4. MLSP Data Plane Actions | 13 |
| 4. Security Considerations | 14 |
| 5. Acknowledgments | 15 |
| 6. IANA Considerations | 16 |
| 7. References | 17 |
| 7.1. Normative References | 17 |
| 7.2. Informative References | 17 |
| Author's Address | 19 |

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

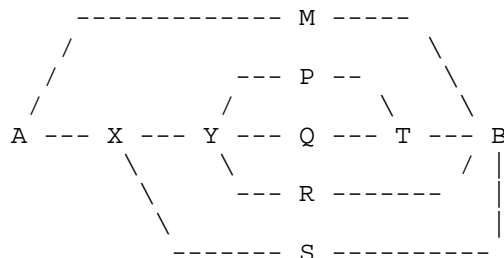
An MLSP is configured at the ingress with various constraints typically associated with TE LSPs, such as destination LSR(s), bandwidth (on a per-class basis, if desired), link colors, Shared Risk Link Groups, etc. [Auto-mesh techniques ([RFC4972]) can be used to reduce configuration; this is not described 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, AXYP_TB, AX_YQ_TB, AX_YR_B and AX_SB 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 AXYP_TB, AX_YQ_TB and AX_YR_B will each carry 10Gbps, and the AX_SB path will carry 30Gbps. We'll call this "IP-style" load balancing.

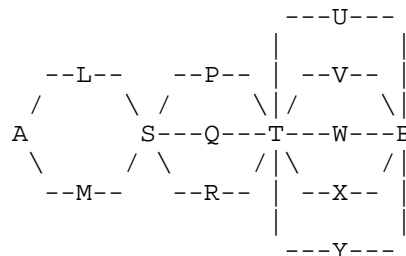
Note: all load balancing is subject to the overriding requirement of mapping the same "flow" to the same downstream. (What constitutes a "flow" is beyond the scope of this document.) This requirement takes precedence over all attempts to balance traffic among downstreams. Thus, the statements above (e.g., "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, AXYP_TB, AX_YQ_TB, AX_YR_B and AX_SB, 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 $2 \times 3 \times 5 = 30$ 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 Y_i 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

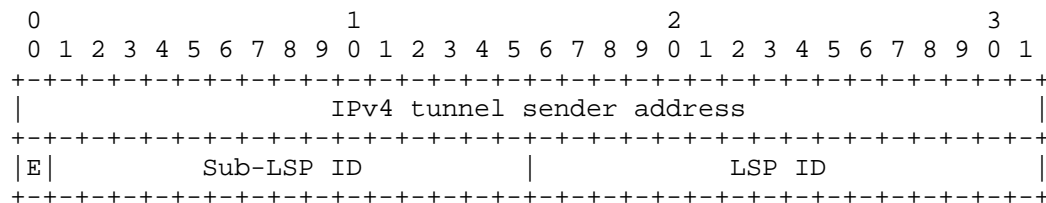


Figure 3: MLSP_TUNNEL_IPv4 Sender Template

Class = SENDER_TEMPLATE, MLSP_TUNNEL_IPv6 C-Type = TBD

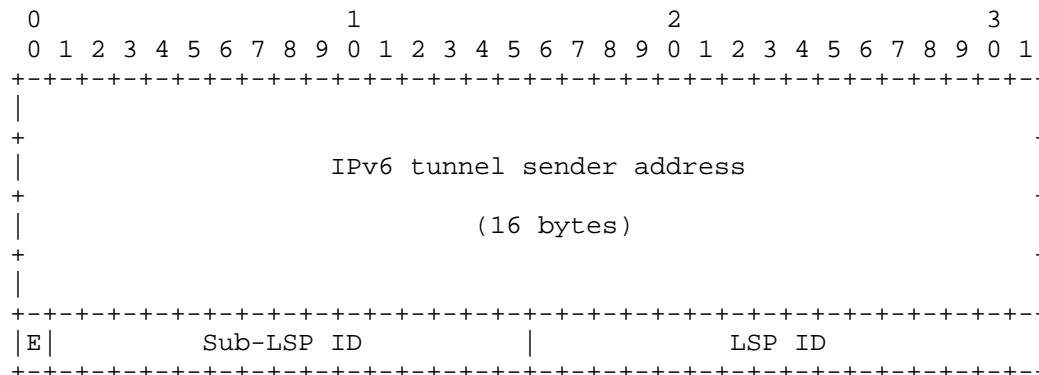


Figure 4: MLSP_TUNNEL_IPv6 Sender Template

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
Contrail Systems
2350 Mission College Blvd.
Santa Clara, CA 94054
US

Email: kireeti.kompella@gmail.com

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

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

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

Abstract

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

Status of this Memo

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

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

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

This Internet-Draft will expire on April 18, 2013.

Copyright Notice

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

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

Table of Contents

| | |
|--|----|
| 1. Introduction | 3 |
| 1.1. Conventions used | 3 |
| 2. Questions | 4 |
| 3. Answers | 5 |
| 3.1. Extended Special Purpose MPLS Label Values | 5 |
| 3.2. Process for Retiring Special Purpose Labels | 6 |
| 4. IANA Considerations | 7 |
| 5. Security Considerations | 8 |
| 6. References | 9 |
| 6.1. Normative References | 9 |
| 6.2. Informational References | 9 |
| Authors' Addresses | 10 |

1. Introduction

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

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

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

1.1. Conventions used

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

2. Questions

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

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

3. Answers

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

1.

- A. Allocation of special purpose MPLS labels is via "Standards Action".
 - B. The IANA registry will be renamed "Special Purpose MPLS Labels".
 - C. Early allocation may be allowed on a case-by-case basis.
 - D. The current space of 16 special purpose labels is too small for setting aside value for experimental or private use. However, the extended special purpose labels registry created by this document has enough space, and this document defines a range for experimental use.
- 2. A Standards Track RFC must accompany a request for allocation of special purpose labels, as per [RFC5226].
 - 3. The retirement of a special purpose MPLS label value must follow a strict and well-documented process. This is necessary since we must avoid orphaning the use of this label value in existing deployments. This process is detailed in Section 3.2.
 - 4. The use of the "implicit null label" (label 3) in the data plane may be allowed, subject to approval by the MPLS WG, and an accompanying Standards Track RFC that details the use of the label, and a discussion of possible sources of confusion between signaling and data plane, and mitigation thereof.
 - 5. The special purpose label (the "extension" label) is to be set aside for the purpose of extending the space of special purpose labels. Further details are described in Section 3.1.

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

3.1. Extended Special Purpose MPLS Label Values

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

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

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

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

3.2. Process for Retiring Special Purpose Labels

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

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

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

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

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

4. IANA Considerations

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

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

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

Table 1

5. Security Considerations

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

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

6. References

6.1. Normative References

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

6.2. Informational References

- [RFC3429] Ohta, H., "Assignment of the 'OAM Alert Label' for Multiprotocol Label Switching Architecture (MPLS) Operation and Maintenance (OAM) Functions", RFC 3429, November 2002.
- [RFC5586] Bocci, M., Vigoureux, M., and S. Bryant, "MPLS Generic Associated Channel", RFC 5586, June 2009.
- [I-D.ietf-mpls-entropy-label]
Kompella, K., Drake, J., Amante, S., Henderickx, W., and L. Yong, "The Use of Entropy Labels in MPLS Forwarding", draft-ietf-mpls-entropy-label-06 (work in progress), September 2012.

Authors' Addresses

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

Email: kireeti.kompella@gmail.com

Loa Andersson
Ericsson

Email: loa@pi.nu

Adrian Farrel
Juniper Networks

Email: adrian@olddog.co.uk

Network Working Group
Internet-Draft
Intended status: Informational
Expires: April 18, 2013

Z. Li
Z. Zhuang
J. Dong
Huawei Technologies
October 15, 2012

A Framework for Service-Driven Co-Routed MPLS Traffic Engineering LSPs
draft-li-mpls-serv-driven-co-lsp-fmwk-00

Abstract

This document provides a framework for setting up service-driven co-routed MPLS Traffic-Engineered Label-Switched Paths (TE LSP) in Multi-Protocol Label Switching (MPLS) networks.

Requirements Language

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

Status of this Memo

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

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

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

This Internet-Draft will expire on April 18, 2013.

Copyright Notice

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

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

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

Table of Contents

| | |
|---|---|
| 1. Introduction | 3 |
| 2. Terminology | 3 |
| 3. Problem Statement | 3 |
| 3.1. MPLS TE Configuration | 3 |
| 3.2. Return Path of BFD for LSP | 3 |
| 3.3. Upgrading of Co-routed Bidirectional LSP | 4 |
| 4. Framework and Procedures | 4 |
| 4.1. Service-Driven Co-Routed Unidirectional LSPs for L2VPN | 4 |
| 4.1.1. Framework | 4 |
| 4.1.2. Procedures | 5 |
| 4.2. Service-Driven Co-Routed Unidirectional LSPs for L3VPN | 6 |
| 4.2.1. Framework | 6 |
| 5. IANA Considerations | 8 |
| 6. Security Considerations | 8 |
| 7. References | 8 |
| 7.1. Normative References | 8 |
| 7.2. Informative References | 8 |
| Authors' Addresses | 9 |

1. Introduction

MPLS Traffic Engineering (TE) has been widely deployed to support packet-based services. Rich Traffic Engineering properties can be provided to satisfy different requirements of services. As the MPLS TE LSP is deployed in networks of service providers, several challenges has been proposed about esay operation and management. This document propose a solution to set up co-routed MPLS TE LSPs on demand to faciliate the deployment for services.

2. Terminology

This document uses terminology from the MPLS architecture document [RFC3031] and from the RSVP-TE protocol specification [RFC3209] which inherits from the RSVP specification [RFC2205].

The PEs can be generally categorized into two types:

1. Active PE: the PE which primarily triggers to set up the LSPs and informs the remote PE;
2. Passive PE: the PE which secondarily complies with the active PE's suggestion.

3. Problem Statement

3.1. MPLS TE Configuration

It is a common deployment scenario to set up MPLS Traffic Engineering(TE) LSPs among a set of Label Switch Routers (LSR). Such deployment may require the configuration of a potentially large number of TE tunnels. The operation is not only time consuming but also prone to misconfiguration for Service Providers. Hence, an automatic mechanism for setting up MPLS TE tunnels is desirable which can simplify the complexity of MPLS TE configuration.

3.2. Return Path of BFD for LSP

BFD for LSP ([RFC5884]) is used to detect the possible failure fast and the failure detected can trigger traffic switch between the primary LSP and the backup LSP. When BFD for LSP is deployed, the return path may take IP path which is different from the forwarding path. The failure that happens in the return path may trigger wrong traffic switch.

3.3. Upgrading of Co-routed Bidirectional LSP

Co-routed bidirectional LSPs can simplify operation and management for Service Providers. Moreover co-routed bidirectional LSPs require less states comparing with two unidirectional MPLS TE LSPs. But the unidirectional LSP has been deployed widely and it is difficult for the service provider to upgrade all possible routers to support co-routed bidirectional LSPs.

4. Framework and Procedures

The section proposes the solution to trigger setting up MPLS TE LSP by services. The framework and procedures of the solution are introduced. With the solution, MPLS TE LSPs can setup on demand which can reduce the statical configuration. In addition, the signalling for the service will advertise the tunnel information between the active PE and the passive PE. The LSP on the passive side can setup according to RRO information of the LSP setup from the active PE to the passive PE. Thus the path of the reverse LSP can be co-routed with the path of the LSP from the active PE to the passive PE.

4.1. Service-Driven Co-Routed Unidirectional LSPs for L2VPN

4.1.1. Framework

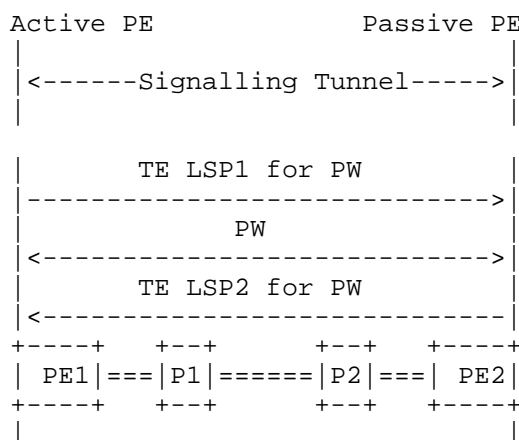


Figure 2: Signalling Tunnel Framework of L2VPN

Figure 2 shows a framework for co-routed MPLS TE LSPs driven by PW. L2VPN is provisioned on PEs and PW is setup. A pair of PEs for a specific PW will be identified as the active PE and the passive PE.

The active PE triggers setting up the primary LSP to the passive PE and advertises the tunnel information to the passive PE. According to the information advertised the passive PE will set up the return MPLS TE LSP which path is co-routed with that of the primary LSP.

4.1.2. Procedures

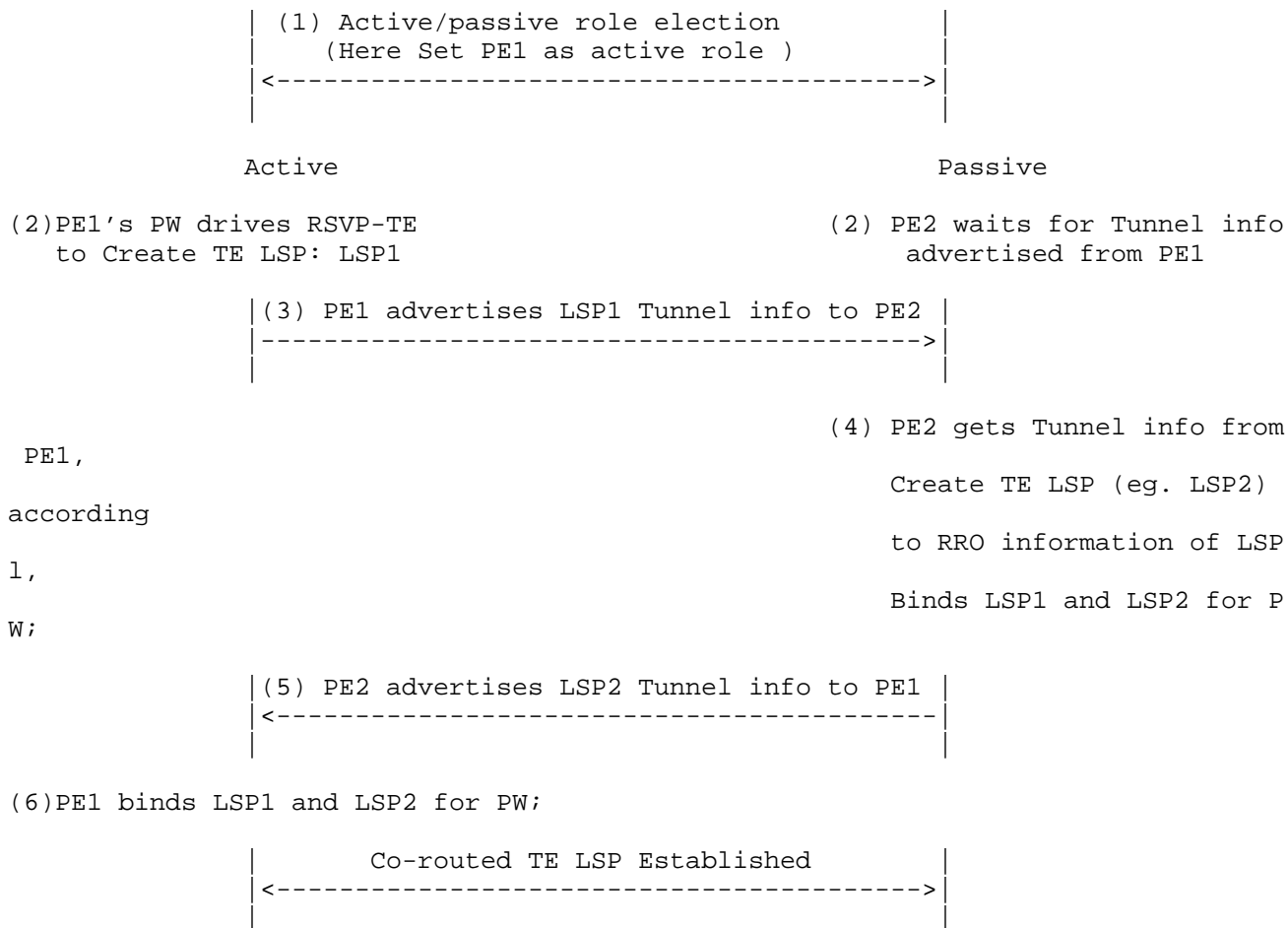


Figure 3: Signalling Procedures of L2VPN

Figure 3 shows the detailed signalling procedures for L2VPN to drive setup of co-routed MPLS TE LSPs:

(1) Active/passive role election through signalling for a pair of PEs of a PW (Assume PE1 as active PE and PE2 as passive PE after election);

(2) PE1's PW drives RSVP-TE to create TE LSP(LSP1) while PE2 waits for tunnel information advertised by PE1;

- (3) PE1 advertises tunnel information related with LSP1 to PE2;
- (4) PE2 gets tunnel information from PE1 and creates TE LSP (eg. LSP2) according to RRO information derived from LSP1. PE2 binds LSP1 and LSP2 for PW;
- (5) PE2 advertises tunnel information related with LSP2 to PE1;
- (6) PE1 binds LSP1 and LSP2 for PW.

Through the above signalling procedure, the co-routed MPLS TE LSPs driven by the PW are established.

4.2. Service-Driven Co-Routed Unidirectional LSPs for L3VPN

4.2.1. Framework

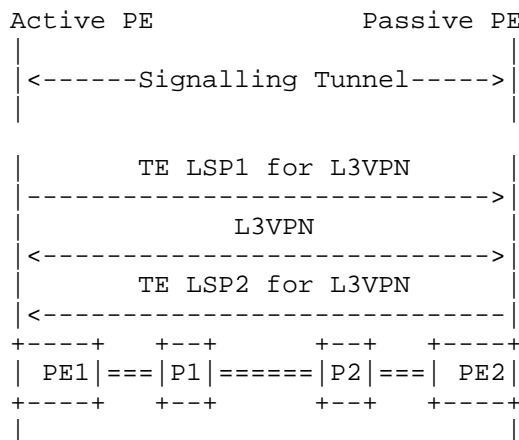


Figure 4: Signalling Tunnel Framework of L3VPN

Figure 4 shows a framework for co-routed MPLS TE LSPs driven by L3VPN. L3VPN is provisioned on PEs and VPN members are discovered. A pair of PEs for L3VPN members will be identified as the active PE and the passive PE. The active PE triggers set up the primary LSP to the passive PE and advertises the tunnel information to the passive PE. According to the information advertised the passive PE will set up the return MPLS TE LSP which path is co-routed with that of the primary LSP.

4.2.1.1. Procedures

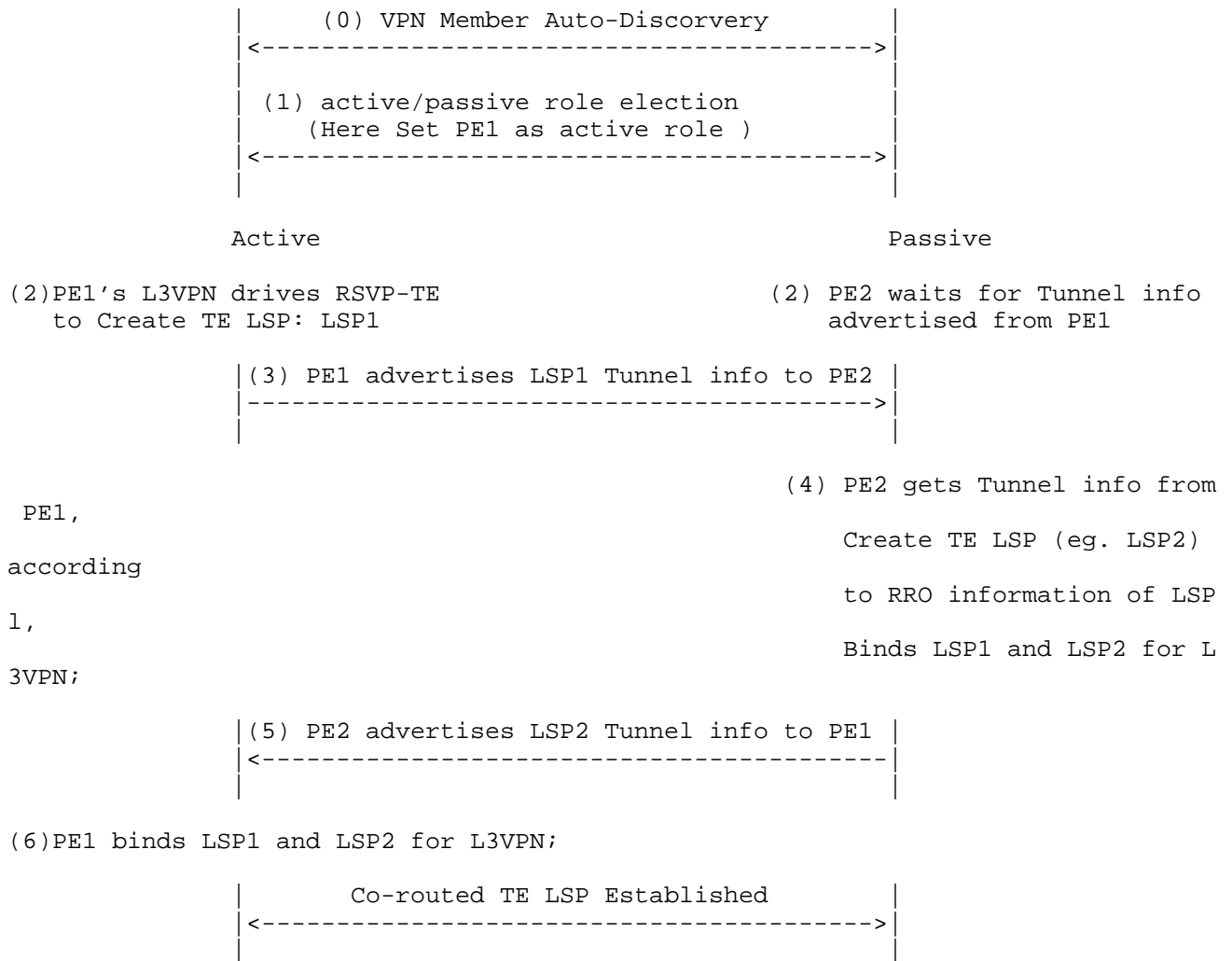


Figure 5: Signalling Procedures of L3VPN

Figure 5 shows the detailed signalling procedures for L3VPN to drive setup of co-routed MPLS TE LSPs :

(0) VPN member auto-discovery process is done through signalling to identify a pair of VPN members;

(1) Active/passive role election through signalling for a pair of PEs of a PW (Assume PE1 as active PE and PE2 as passive PE after election);

(2) PE1's PW drives RSVP-TE to create TE LSP(LSP1) while PE2 waits for tunnel information advertised by PE1;

(3) PE1 advertises tunnel information related with LSP1 to PE2;

(4) PE2 gets tunnel information from PE1 and creates TE LSP (eg. LSP2) according to RRO information derived from LSP1. PE2 binds LSP1 and LSP2 for L3VPN;

(5) PE2 advertises tunnel information related with LSP2 to PE1;

(6) PE1 binds LSP1 and LSP2 for L3VPN.

Through the above signalling procedure, the co-routed MPLS TE LSPs driven by the L3VPN are established.

5. IANA Considerations

This document makes no request of IANA.

6. Security Considerations

This document does not change the security properties of L2VPN & L3VPN.

7. References

7.1. Normative References

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

7.2. Informative References

[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.

[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

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

Email: lizhenbin@huawei.com

Shunwan Zhuang
Huawei Technologies
Huawei Bld., No.156 Beiqing Rd.
Beijing 100095
China

Email: zhuangshunwan@huawei.com
URI: jie.dong@huawei.com

Jie Dong
Huawei Technologies
Huawei Bld., No.156 Beiqing Rd.
Beijing 100095
China

Email: jie.dong@huawei.com

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

G. Swallow
V. Lim
Cisco Systems
October 12, 2012

Proxy LSP Ping
draft-lim-mpls-proxy-lsp-ping-00

Abstract

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

Status of this Memo

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

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

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

This Internet-Draft will expire on April 15, 2013.

Copyright Notice

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

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

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

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

Table of Contents

| | |
|---|----|
| 1. Introduction | 3 |
| 1.1. Requirements Language | 3 |
| 2. Proxy Ping Overview | 4 |
| 3. Proxy MPLS Echo Request / Reply Procedures | 5 |
| 3.1. Procedures for the initiator | 5 |
| 3.2. Procedures for the proxy LSR | 7 |
| 3.2.1. Downstream Detailed/Downstream Maps in Proxy Reply . . | 8 |
| 3.2.2. Sending an MPLS proxy ping reply | 9 |
| 3.2.3. Sending the MPLS echo requests | 9 |
| 3.2.3.1. Forming the base MPLS echo request | 9 |
| 3.2.3.2. Per interface sending procedures | 10 |
| 4. Proxy Ping Request / Reply Messages | 11 |
| 4.1. Proxy Ping Request / Reply Message formats | 11 |
| 4.2. Proxy Ping Request Message contents | 12 |
| 4.3. Proxy Ping Reply Message Contents | 12 |
| 5. Object formats | 12 |
| 5.1. Proxy Echo Parameters Object | 12 |
| 5.1.1. Next Hop sub-Object | 15 |
| 5.2. Reply-to Address Object | 16 |
| 5.3. Upstream Neighbor Address Object | 17 |
| 5.4. Downstream Neighbor Address Object | 18 |
| 6. Security Considerations | 19 |
| 7. IANA Considerations | 20 |
| 8. References | 21 |
| 8.1. Normative References | 21 |
| 8.2. Informative References | 21 |
| Authors' Addresses | 21 |

1. Introduction

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

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

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

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

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

1.1. Requirements Language

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

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

ignored on receipt.

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

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

2. Proxy Ping Overview

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

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

We refer to these nodes with the following terms:

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

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

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

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

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

The initiator formats an MPLS proxy ping request message and sends it to the proxy LSR, a node it believes to be on the path of the LSP. This message specifies the MPLS echo request to be sent inband of the LSP. It may request the proxy LSR to either Reply with Proxy information or the send a MPLS echo request. The initiator requests Proxy information so that it can learn additional information it needs to use to form a subsequent MPLS Proxy Ping request. For example during LSP traceroute an initiator needs the downstream map

information to form an Echo request. An initiator may also want to learn a Proxy LSR's FEC neighbor information so that it can form proxy request to various nodes along the LSP.

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

If the proxy LSR has a label mapping for the FEC and all authorization checks have passed, the proxy LSR formats an MPLS echo request. If the source address of the IP packet is not the initiator, it includes a Reply-to Address object containing the initiator's address. It then sends it inband of the LSP.

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

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

3. Proxy MPLS Echo Request / Reply Procedures

3.1. Procedures for the initiator

The initiator creates an MPLS proxy ping request message.

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

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

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

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

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

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

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

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

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

Pad

Vendor Enterprise Number

Reply TOS Byte

P2MP Egress Identifier [RFC6425]

Echo Jitter TLV [RFC6425]

Vendor Private TLVs

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

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

IP TTL is set to 255.

3.2. Procedures for the proxy LSR

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

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

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

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

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

The destination address specified in the Proxy Echo Parameters object is checked to ensure that it conforms to the address allowed IPv4 or

IPv6 address range. If not, it sets the Return Code set to "Malformed echo request received" and the Subcode set to zero. If the Reply Mode of the message header is not 1, an MPLS proxy ping reply message SHOULD be sent as described below.

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

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

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

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

If the Reply Mode of the message header is 1, 2 or 3 and no errors or modifications have occurred no MPLS proxy ping reply is sent. Otherwise an MPLS proxy ping reply message SHOULD be sent as described below.

3.2.1. Downstream Detailed/Downstream Maps in Proxy Reply

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

When the Proxy LSR is a bud node or egress in a MP2MP LSP, the Proxy Reply should contain DMAP/DDMAPs assuming that the packet is being sourced from a leaf. In this case, there will be no DMAP/DDMAP

describing the egresses. The Proxy reply return code is either set to "Reply router found mapping for the FEC" or "Reply router is Egress for the FEC" is returned.

3.2.2. Sending an MPLS proxy ping reply

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

3.2.3. Sending the MPLS echo requests

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

3.2.3.1. Forming the base MPLS echo request

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

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

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

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

A Reply-to Address object containing the initiator's address is included.

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

Target FEC Stack

Pad

Vendor Enterprise Number

Reply TOS Byte

P2MP Egress Identifier [RFC6425]

Echo Jitter TLV [RFC6425]

Vendor Private TLVs

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

The source IP address is set to a routable address of the proxy LSR. Per usual the TTL of the IP packet is set to 1.

If the Explicit DSCP flag is set, the Requested DSCP byte is examined. If the setting is permitted then the DSCP byte of the IP header of the MPLS Echo Request message is set to that value. Otherwise the DSCP byte is set to a default value. In this case the MPLS Proxy Echo Parameters with the Explicit DSCP flag cleared MUST be included in any MPLS proxy ping reply message. The return code MUST be set to <tba>, "Proxy ping parameters modified". The DSCP field of the MPLS Proxy Echo Parameters SHOULD be set to the actual value used.

3.2.3.2. Per interface sending procedures

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

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

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

The packet is then transmitted on this interface.

4. Proxy Ping Request / Reply Messages

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

4.1. Proxy Ping Request / Reply Message formats

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

```

      0               1               2               3
      0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
+-----+-----+-----+-----+-----+-----+-----+-----+
|                                     |                                     |
|          Version Number           |          MUST Be Zero             |
+-----+-----+-----+-----+-----+-----+-----+-----+
|  Message Type  |  Reply mode  |  Return Code  |  Return Subcode  |
+-----+-----+-----+-----+-----+-----+-----+-----+
|                                     |                                     |
|                                     |          Sender's Handle          |
+-----+-----+-----+-----+-----+-----+-----+-----+
|                                     |          Sequence Number          |
+-----+-----+-----+-----+-----+-----+-----+-----+
|                                     |          TLVs ...                  |
|                                     |                                     |
|                                     |                                     |
|                                     |                                     |
+-----+-----+-----+-----+-----+-----+-----+-----+

```

Version Number

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

Message Type

| Type | Message |
|------|--|
| ---- | ----- |
| 5 | MPLS proxy ping request
(Pending IANA assignment) |
| 6 | MPLS proxy ping reply
(Pending IANA assignment) |

4.2. Proxy Ping Request Message contents

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

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

4.3. Proxy Ping Reply Message Contents

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

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

5. Object formats

5.1. Proxy Echo Parameters Object

The Proxy Echo Parameters object is a TLV that MUST be included in an MPLS Proxy Echo Request message. The length of the TLV is $12 + K + S$, where K is the length of the Destination IP Address field and S is the total length of the sub-objects. The Proxy Echo Parameters object can be used to either to 1) control attributes used in Composing and Sending an MPLS echo request or 2) query the Proxy LSR for information about the topmost FEC in the target FEC stack but not

both. In the case where the Proxy LSR is being queried (ie information needs to be returned in a Proxy Reply), no MPLS echo request will be sent from the Proxy LSR. The MPLS Proxy Echo request echo header's Reply Mode should be set to "Reply with Proxy Info".

| 0 | | | | | | | | | | 1 | | | | | | | | | | 2 | | | | | | | | | | 3 | | | | | | | | | |
|------------------------|---|---|---|---|---|---|---|---|---|-------------------|---|---|---|---|---|---|---|---|---|-----------------|---|---|---|---|---|---|---|---|---|---|---|--|--|--|--|--|--|--|--|
| 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 0 | 1 | | | | | | | | |
| Address Type | | | | | | | | | | Reply mode | | | | | | | | | | Proxy Flags | | | | | | | | | | | | | | | | | | | |
| TTL | | | | | | | | | | Rqst'd DSCP | | | | | | | | | | Source UDP Port | | | | | | | | | | | | | | | | | | | |
| Global Flags | | | | | | | | | | MPLS Payload size | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| Destination IP Address | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| Sub-Objects | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |

Address Type

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

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

Reply mode

The reply mode to be sent in the MPLS Echo Request message; the values are as specified in [RFC4379].

Proxy Flags

Request for FEC Neighbor Address info 0x01

When set this requests that the proxy LSR supply the Upstream and Downstream neighbor address information in the MPLS proxy ping reply message. This flag is only applicable

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

Setting this flag will cause the proxy LSR to cancel sending an Echo request as the information being requested needs to be returned for use in a subsequent Proxy Request.

Request for Downstream Mapping 0x02

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

Setting this flag will cause the proxy LSR to cancel sending an Echo request as the information being requested needs to be returned for use in a subsequent Proxy Request.

Request for Enhanced Downstream Mapping 0x04

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

Setting this flag will cause the proxy LSR to cancel sending an Echo request as the information being requested needs to be returned for use in a subsequent Proxy Request.

Explicit DSCP Request 0x08

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

TTL

The TTL to be used in the label stack entry corresponding to the topmost FEC in the in the MPLS Echo Request packet

Requested DSCP

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

Source UDP Port

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

Global Flags

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

MPLS Payload Size

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

Destination IP Address

If the Address Type is IPv4, an address from the range 127/8;
If the Address Type is IPv6, an address from the range
::FFFF:7F00:0/104

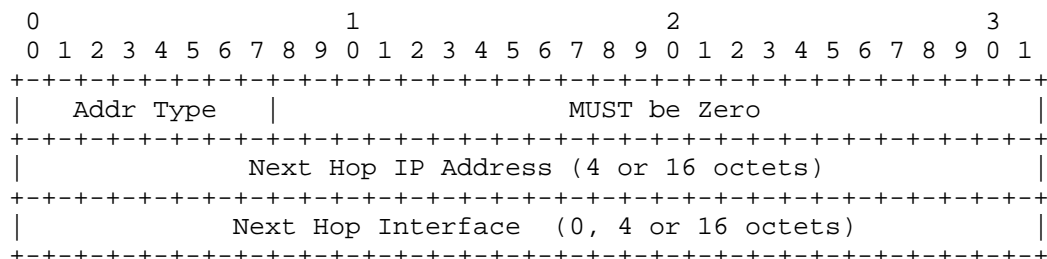
Sub-Objects

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

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

5.1.1. Next Hop sub-Object

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



Address Type

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

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

Next Hop IP Address

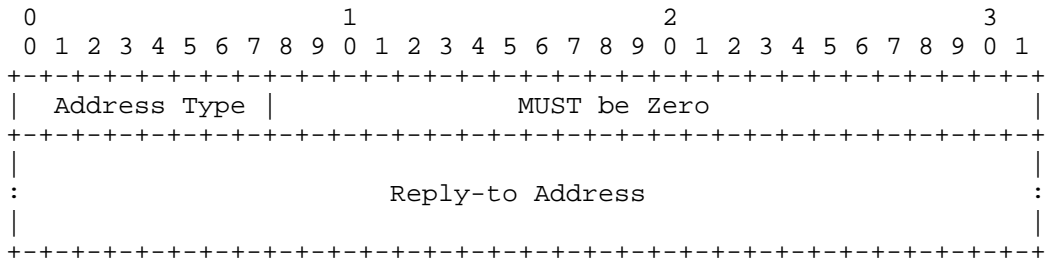
A next hop address that the echo request message is to be sent towards

Next Hop Interface

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

5.2. Reply-to Address Object

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

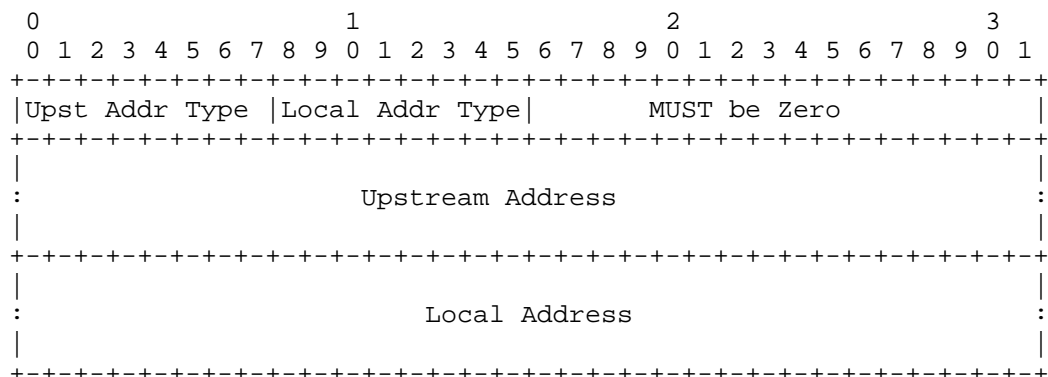


Address Type

A type code as specified in the table below:

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

5.3. Upstream Neighbor Address Object



Upst Addr Type; Local Addr Type

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

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

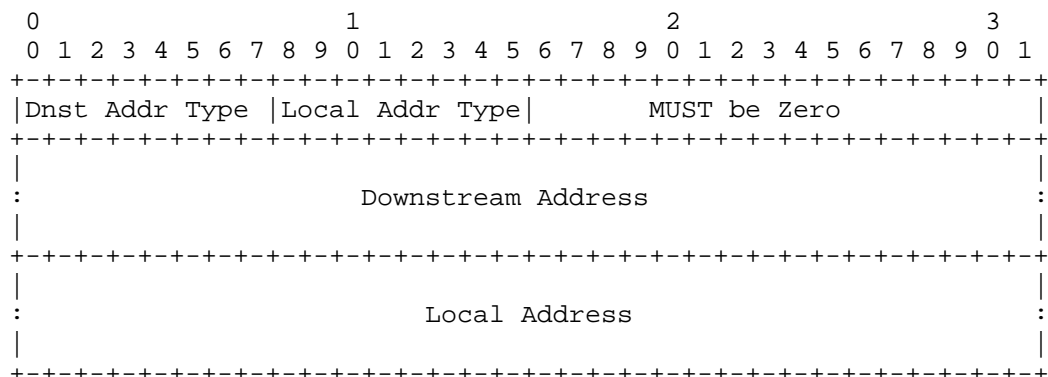
Upstream Address

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

Local Address

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

5.4. Downstream Neighbor Address Object



Upst Addr Type; Local Addr Type

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

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

Upstream Address

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

Local Address

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

6. Security Considerations

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

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

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

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

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

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

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

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

7. IANA Considerations

This document makes the following assignments (pending IANA action)

LSP Ping Message Types

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

Objects and Sub-Objects

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

Return Code [pending IANA assignment]

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

8. References

8.1. Normative References

- [RFC4379] Kompella, K. and G. Swallow, "Detecting Multi-Protocol Label Switched (MPLS) Data Plane Failures", RFC 4379, February 2006.
- [RFC6424] Bahadur, N., Kompella, K., and G. Swallow, "Mechanism for Performing Label Switched Path Ping (LSP Ping) over MPLS Tunnels", RFC 6424, November 2011.
- [RFC6425] Saxena, S., Swallow, G., Ali, Z., Farrel, A., Yasukawa, S., and T. Nadeau, "Detecting Data-Plane Failures in Point-to-Multipoint MPLS - Extensions to LSP Ping", RFC 6425, November 2011.

8.2. Informative References

- [RFC4875] Aggarwal, R., Papadimitriou, D., and S. Yasukawa, "Extensions to Resource Reservation Protocol - Traffic Engineering (RSVP-TE) for Point-to-Multipoint TE Label Switched Paths (LSPs)", RFC 4875, May 2007.
- [RFC6388] Wijnands, IJ., Minei, I., Kompella, K., and B. Thomas, "Label Distribution Protocol Extensions for Point-to-Multipoint and Multipoint-to-Multipoint Label Switched Paths", RFC 6388, November 2011.

Authors' Addresses

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

Email: swallow@cisco.com

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

Email: vlim@cisco.com

MPLS Working Group
Internet-Draft
Intended status: Informational
Expires: April 22, 2013

G. Liu
ZTE Corporation
Y. Weigarten

M. Daikoku
T. Maruyama
KDDI Corporation
October 19, 2012

MPLS-TP protection for interconnected rings
draft-liu-mpls-tp-interconnected-ring-protection-03

Abstract

The requirements for MPLS Transport Profile include a requirement (R93) that requires that MPLS-TP must support recovery mechanisms for a network constructed from interconnected rings that protect user data that traverses more than one ring. In particular, this includes protecting against cases of failure at the ring interconnect nodes and links. This document presents different configurations of interconnected rings and special mechanisms to address the recovery of ring-interconnect nodes and links. .

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

Status of this Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79. This document may not be modified, and derivative works of it may not be created, and it may not be published except as an Internet-Draft.

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

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

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

Copyright Notice

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

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

Table of Contents

| | |
|---|----|
| 1. Introduction | 4 |
| 2. Conventions used in this document | 8 |
| 3. recovery mechanism | 9 |
| 3.1. recovery mechanism for Dual-node interconnection | 9 |
| 3.2. recovery mechanism for Chained interconnection | 11 |
| 4. Security Considerations | 12 |
| 5. IANA Considerations | 12 |
| 6. Acknowledgments | 12 |
| 7. References | 12 |
| 7.1. Normative References | 12 |
| 7.2. Informative References | 12 |
| 7.3. URL References | 12 |
| Authors' Addresses | 13 |

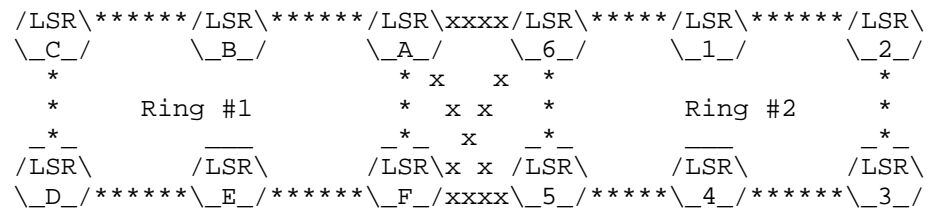
1. Introduction

This document describes different interconnected ring scenarios and a few special solutions to protect against the failure of the ring-interconnect nodes and links. there are three common interconnection scenarios that we will address in this document:

Dual-node interconnection - when the two rings are interconnected by two nodes from each ring (see Figure 1);

Single-node interconnection - when the connection between the two rings is through a single node (see Figure 2).As the interconnection node(LSR-A) is a single-point of failure, This configuration should be avoided in real networks;

Chained interconnection - 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

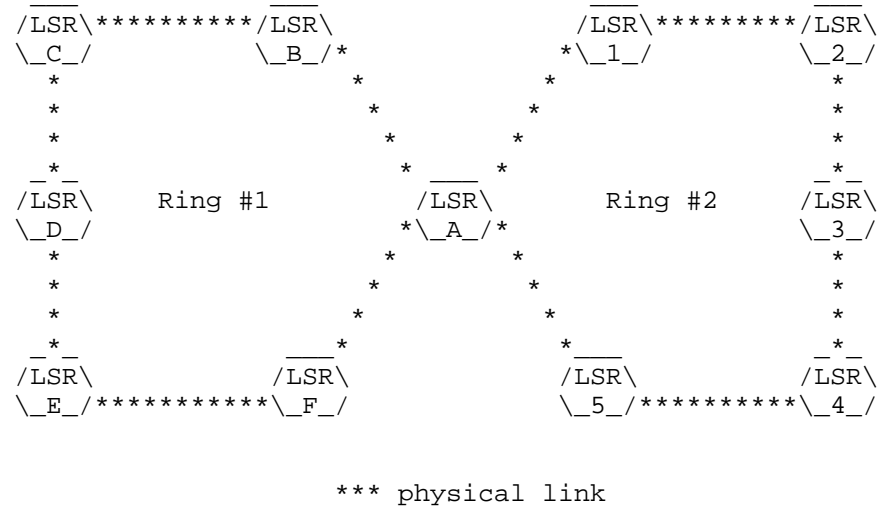


Figure 2

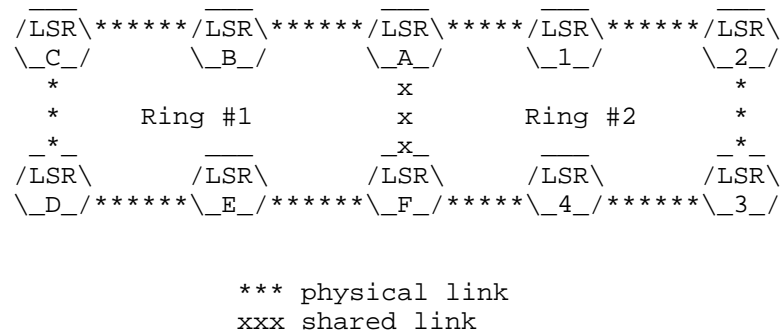


Figure 3

Regarding traffic that traverses more than two rings. many interconnection scenarios could be existed in the same scenario, they will be mixed interconnection scenario:

Dual-node and single-node mixed interconnection- when there exist a multi-ring traffic which traverses more than two ring. two of these rings are dual-node interconnection. while another two are single-node interconnection (see figure 5);

Dual-node and chained mixed interconnection-when there exist both dual-node interconnection and chained interconnection in this scenario (see figure 4);

single-node and chained mixed interconnection-when there exist both single-node interconnection and chained interconnection in this scenario(see figure 6);

Dual-node, single-node and chained mixed interconnection-when there exist all three interconnection scenarios in this scenario including Dual-node interconnection, single-node interconnection and chained interconnection(see figure 7);

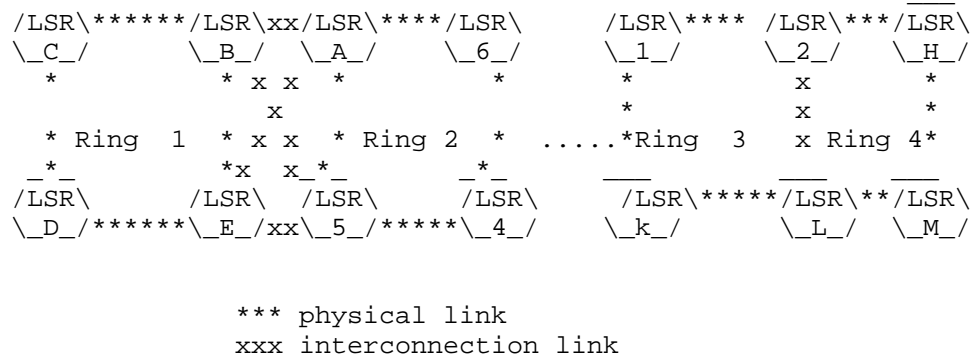


Figure 4

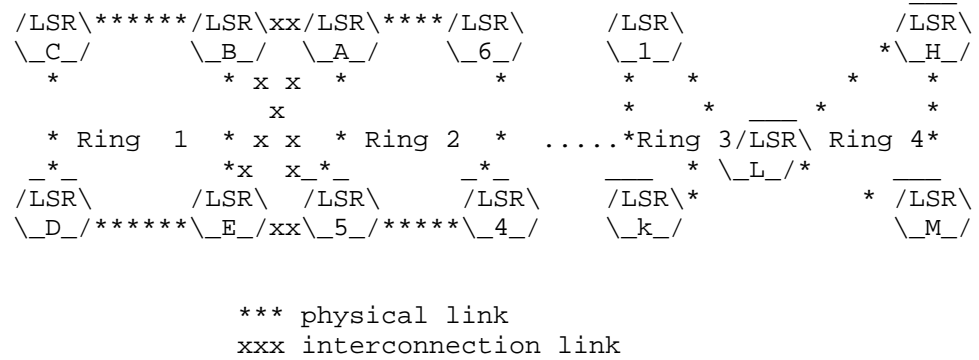


Figure 5

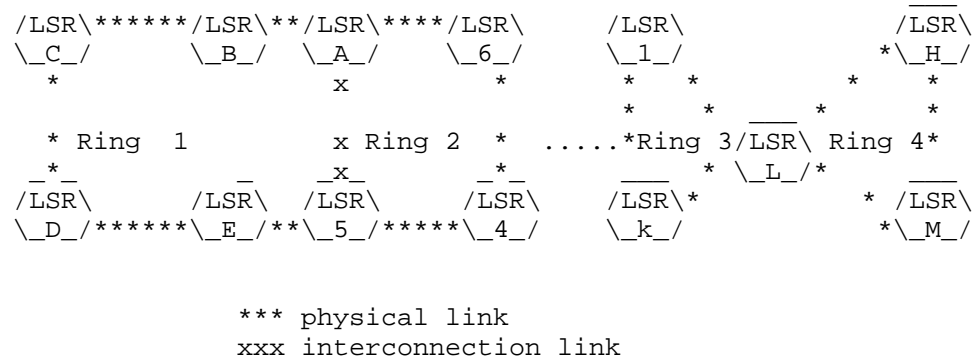


Figure 6

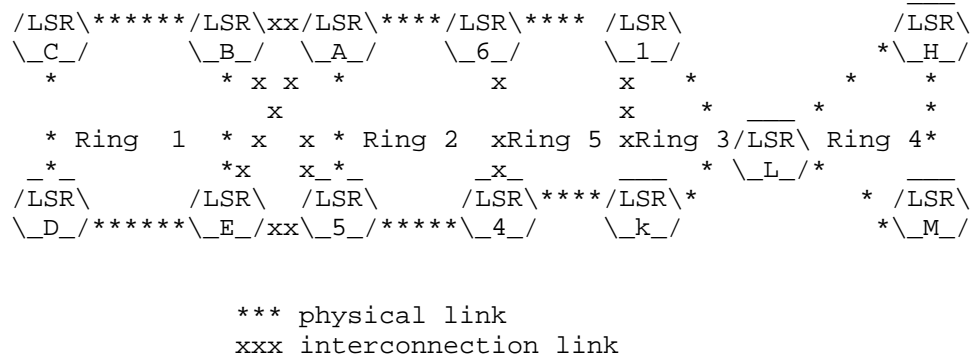


Figure 7

For a multi-ring traffic, it will be across more than one ring just like above seven scenarios. if a failure happens on a multi-ring path, quick recovery is necessary requirement for multi-ring traffic.

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

PSC:Protection Switching Coordination

SD:Signal Degrade

SF:Signal Fail

MPLS-TP:Multi-Protocol Label Switching Transport Profile

3. recovery mechanism

In the following subsections we propose different mechanisms that may be applied for traffic recovery in the different interconnection scenarios. In general, it is possible to provide protection against the failure of a ring node/link by using the single-ring protection mechanism. These cases are out of scope for this document. It is also possible to configure an end-to-end protection to protect the entire working path across all of the interconnected rings. However, this protection scheme does not scale very well. Therefore, we need to consider special mechanisms to address recovery from failures of the interconnecting nodes and links

3.1. recovery mechanism for Dual-node interconnection

Under this scenario , when interconnection link(LSRA-LSR6) has a failure as shown in figure 8. it is possible use 1:1 linear protection mechanism to protect the failure of segment(LSRA-LSR6) by using one of the protection paths (LSRA-LSRF-LSR5-LSR6 or LSRA-LSRF-LSR6 or LSRA-LSR5-LSR6) .

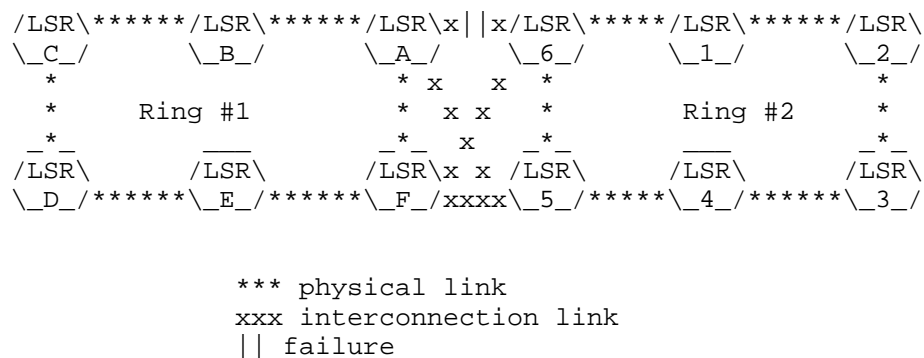


Figure 8

When the interconnection node(LSRA or LSR6) detects a SF or SD on the interconnection link(LSRA-LSR6), LSRA or LSR6 will send SF or SD failure message to its peer node. Then it switches the multi-ring traffic from the working path to its corresponding protection path to another end point(LSRA or LSR6) of the segment . when the peer node (LSR6 or LSRA) receives the traffic packet from its protection

protection path, it will POP the outer label of protection tunnel and return back to the original working tunnel(LSRA-LSRB-LSRC or LSR6-LSR1-LSR2) of another ring(ring 1 or ring 2) to transport the multi-ring traffic.

when interconnection node(LSRA or LSR6) has a failure as shown in figure 9. the end node of the segment detects the failure of the interconnection node, it should send failure message to the backup interconnection node(LSRF or LSR5) to active the protection path that goes to the backup interconnection node(LSRF or LSR5) to transport the multi-ring traffic. at the same time, the backup interconnection node should active its corresponding protection path that goes to another primary interconnection node(LSR6 or LSRA).Then the multi-ring traffic should return back to the original working path to be transported in another primary interconnection node..

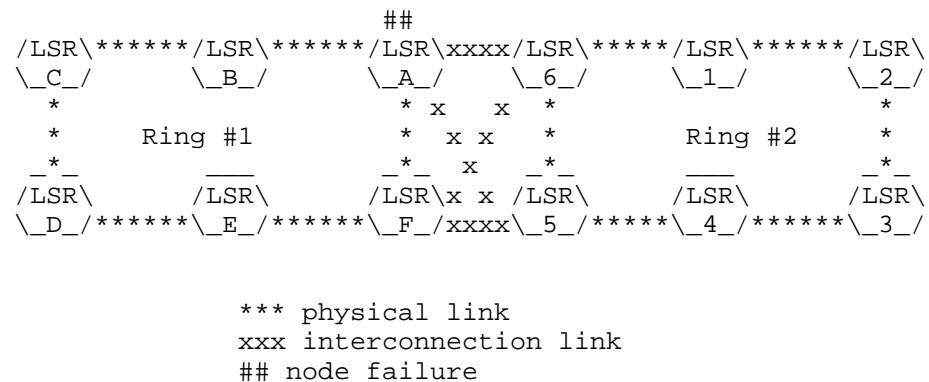


Figure 9

for example , LSRC detects a failure on the interconnection node LSRA. it will send the failure message to notify the backup interconnection node LSRF to switch over to the protection path(LSRC-LSRD-LSRE-LSRF) to transport the multi-ring traffic.at the same time, LSRF should active its corresponding protection path that goes to another primary interconnection node LSR6 to transport the multi-ring traffic.The corresponding protection path may be one of the two paths(LSRF-LSR5-LSR6 or LSRF-LSR6). Then the multi-ring traffic will be transported by its original working path(LSR6-LSR1-LSR2) to the exit node LSR2.

3.2. recovery mechanism for Chained interconnection

For this scenario , when only a failure is detected on the interconnection link. Since the failure should not affect the multi-ring traffic. no action is taken. when a failure happens on the segment of the multi-ring path just as shown in figure 10. The end node of the segment detects the failures, it will active the protection path that goes to the backup interconnection node to transport the multi-ring traffic. After the backup interconnection node receives the failure message , it will active its corresponding protection path that goes to the exit node of another ring to transport the multi-ring traffic.

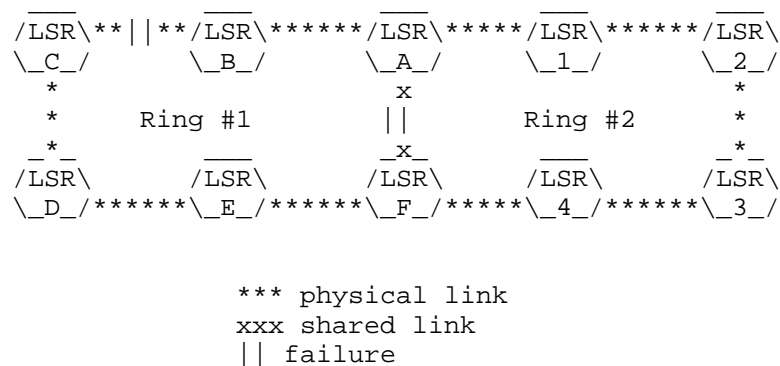


Figure 10

for example, there are failures on both link(LSRC-LSRB) and (LSRA-LSRF) at the same time as shown in figure.10. when LSRC detects or is notified of the failures on both the segment of the working path and the interconnection link. so it will send a failure message to the backup interconnection node LSRF, Then LSRF will active its corresponding protection path(LSRF-LSR4-LSR3-LSR2) of ring 2 to transport the multi-ring traffic.

(Editor's note:should supply text that describes protection against the failure of interconnection node in the chained interconnection

scenario in the future. welcome all experts provide good solution for the 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.

[RFC 6378]
S. Bryant, N. Sprecher, A. Fulignoli Y. Weingarten, "MPLS transport profile Linear Protection", September 2011.

7.2. Informative References

[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

Guoman Liu
ZTE Corporation
No.50, Ruanjian Ave, Yuhuatai District
Nanjing 210012
P.R.China

Phone: +86 025 88014227
Email: liu.guoman@zte.com.cn

Yaacov Weingarten
34 Hagefen St Karnei
Shomron 44853
Israel

Phone: +972-9-775 1827
Email: wyaacov@gmail.com

Masahiro Daikoku
KDDI Corporation
Garden Air Tower, Iidabashi, Chiyoda-ku
Tokyo 102-8460
Japan

Email: ms-daikoku@kddi.com

Takeshi Maruyama
KDDI Corporation
Garden Air Tower, Iidabashi, Chiyoda-ku
Tokyo 102-8460
Japan

Email: ta-maruyama@kddi.com

MPLS Working Group
Internet-Draft
Intended status: Informational
Expires: March 15, 2013

G. Liu
ZTE Corporation
September 11, 2012

Applicability of MPLS-TP Linear protection for p2mp
draft-liu-mpls-tp-p2mp-linear-protection-00

Abstract

In MPLS-TP requirement document(rfc5654), there is a requirement to support 1+1 and 1:n linear protection for p2mp connectivity. The requirement was described in MPLS-TP Survivability Framework document(RFC 6372). The basic protocol for linear protection was specified in the MPLS-TP Linear Protection document [RFC 6378] but is limited to 1+1 and 1:1 protection for p2p connectivity. In addition, The 1:N protection in which all of working transport paths and the protection path have the same end points was specified in MPLS-TP 1:N protection document(draft-ietf-mpls-tp-ltoN-protection). This document applies the existing PSC(RFC 6378) and extensive PSC protocol(draft-ietf-mpls-tp-ltoN-protection) to support scenarios of protecting the p2mp path by extending the existing p2p linear protection mechanism. .

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

Status of this Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79. This document may not be modified, and derivative works of it may not be created, and it may not be published except as an Internet-Draft.

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

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

material or to cite them other than as "work in progress."

This Internet-Draft will expire on March 15, 2013.

Copyright Notice

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

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

Table of Contents

| | |
|--|----|
| 1. Introduction | 4 |
| 1.1. 1+1 p2mp protection | 4 |
| 1.2. 1:1 p2mp per-tree protection | 5 |
| 1.3. 1:1 p2mp per-leaf protection | 6 |
| 1.4. 1:1 p2mp branch path protection | 6 |
| 1.5. 1:n p2mp shared protection | 7 |
| 2. Conventions used in this document | 9 |
| 3. Coordination protocol | 9 |
| 4. switch operation | 10 |
| 4.1. 1+1 protection operation | 10 |
| 4.2. 1:1 p2mp per-tree protection operation | 10 |
| 4.3. 1:1 p2mp per-leaf protection operation | 11 |
| 4.4. 1:1 p2mp branch path protection operation | 11 |
| 4.5. 1:n p2mp shared protection operation | 11 |
| 5. Security Considerations | 12 |
| 6. IANA Considerations | 12 |
| 7. Acknowledgments | 12 |
| 8. References | 12 |
| 8.1. Normative References | 12 |
| 8.2. Informative References | 12 |
| 8.3. URL References | 13 |
| Author's Address | 13 |

1. Introduction

The MPLS Transport Profile(MPLS-TP) Requirement document(RFC5654) and MPLS-TP P2MP Framework document both describe the requirement that unidirectional 1+1 and 1:n protection for p2mp connectivity must be supported. while the MPLS-TP Survivability Framework(RFC6372) is a framework for survivability in MPLS-TP network.

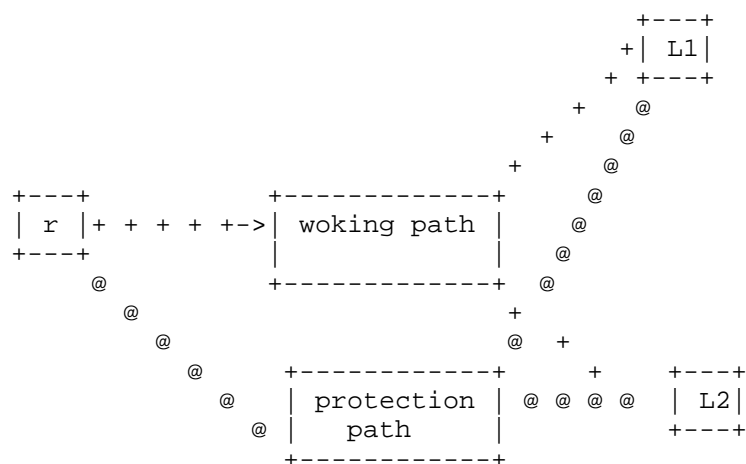
MPLS-TP Linear protection document(RFC6378) defines a Protection State Coordination(PSC) protocol that supports the different 1+1 and 1:1 architectures described in MPLS-TP Survivability Framework. The PSC protocol is a single-phase protocol that allows the two endpoints of the protection domain to coordinate the protection switching operation domain to coordinate the protection switching operation when a switching condition is detected on the transport paths of the protection domain.

MPLS-TP 1:N protection document(draft-ietf-mpls-tp-ltoN-protection) uses a two-phase extensive PSC protocol to protect multiple working paths by a single protection path.

As for the p2mp path, it has multiple leaf nodes and is still unidirectional. so it is not good for p2mp path to use the protection mechanism in RFC6378 and draft-ietf-mpls-tp-ltoN-protection. the document is an applicability of the existing PSC protocol and the extensive PSC protocol to support protection of p2mp path by extending the existing linear protection mechanism.

1.1. 1+1 p2mp protection

This protection is a specific protection that use one designated protection path to protect one working path. It is fully allocated in the sense that the route and bandwidth resource of the protection path is reserved for the working path.under any condition, the source root node will bridge both the working path and the protection path at the same time. while the sink leaf nodes will select one of the two paths to receive the traffic . As it is unnecessary to coordinate the switch state between the root node and the leaf nodes , the PSC protocol is not needed in the 1+1 protection mechanism.



NOTE :

```
@@@@@: p2mp protection path
```

```
+++++: p2mp working path
```

Figure 1

Figure 1 shows a protection domain with one working path and its corresponding protection path. for 1+1 protection, the protection path must transport the p2mp traffic at any time. so the source root node always bridges the p2mp traffic into both the working path and the protection path. while the sink leaf node L1 and L2 will select one of the two paths to receive the traffic based on the performance of the two paths. As it is no sense to coordinate the switch state between the source root node and the sink leaf nodes. it can't apply the PSC protocol for the protection mechanism;

1.2. 1:1 p2mp per-tree protection

The protection will still use one designated p2mp protection path to protect one p2mp working path. but the source root node only bridges one of the two paths at any time. All sink leaf nodes select the same p2mp path to receive the traffic. in addition, As the root node

and leaf nodes need to coordinate to select the same p2mp path to send and receive the p2mp traffic. it may apply the existing PSC protocol to coordinate the switch state between the root node and the leaf nodes.

just as the above figure 1, under normal condition, the traffic will be transported by only the working path from node r to node L1 and L2. when a defect is detected on the working path. the source node r will bridge the traffic into the p2mp protection path. then it should send PSC packet to all leaf nodes L1 and L2 in order to coordinate to select the p2mp protection path to send and receive the p2mp traffic.

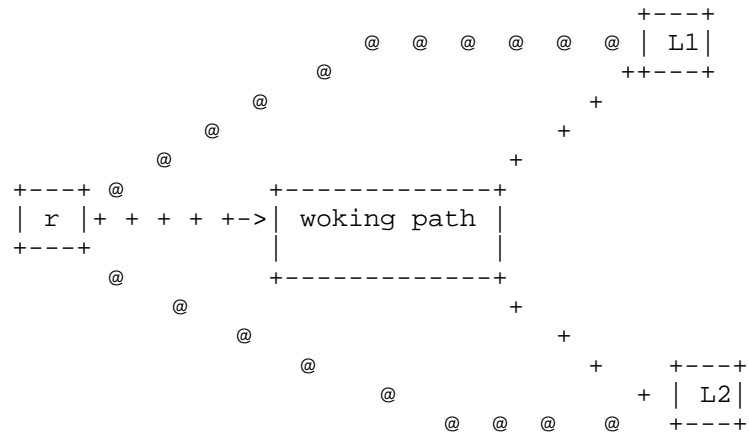
1.3. 1:1 p2mp per-leaf protection

The protection is similar to the above protection mechanism. it need to pre-configure one p2mp protection path to protect the p2mp working path. when a defect is detected on the working path. The source root node firstly bridge both the working path and the protection path to send the traffic on both paths. while the leaf nodes will select one of the two paths to receive the traffic based on whether the working path has defect. As the source root node bridge two paths after a defect is detected on the working path, it is unnecessary to use PSC protocol to coordinate the switch state between the root node and the leaf nodes.

just as the above figure 1, when a defect is detected on one branch path of the p2mp working path, for example, path(r-L1) or path(r-L2) has a defect, The source root node r will firstly send the traffic on both the working path and the protection path. L1 or L2 will select one of the two path to receive the traffic based on whether to have a defect on its branch path(r-L1) or path(r-L2)..

1.4. 1:1 p2mp branch path protection

The protection will pre-configure one p2p protection path to protect each branch path of the p2mp working path. and each branch path is disjointed from its p2p protection path. the protection mechanism is the same as 1:1 protection in MPLS-TP Linear protection(RFC3678), and it needs to use the PSC protocol to coordinate the switch state between the root node and the leaf nodes.



NOTE:

@@@@@: p2p protection path

+++++: p2mp working path

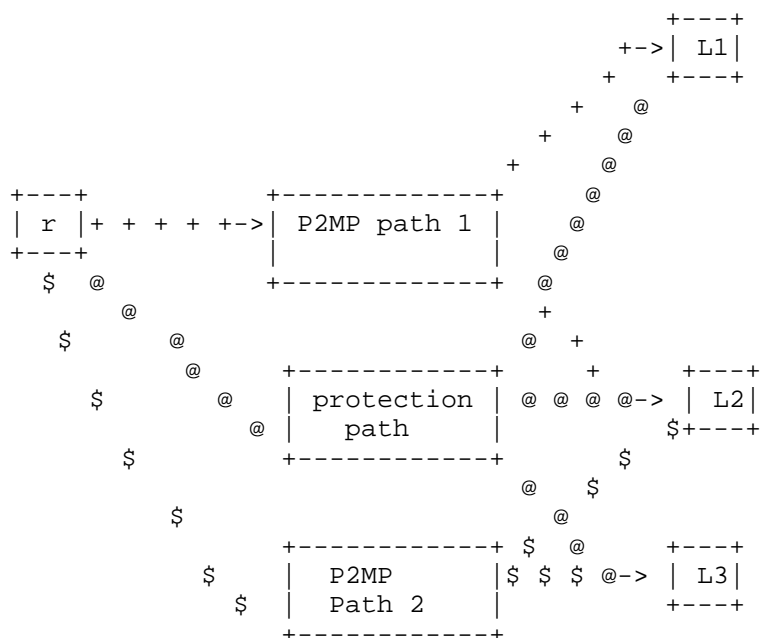
Figure 2

just as the above figure 2, firstly it pre-configures one p2p protection path for each branch path(r-L1) and (r-L2) individually. when a defect is detected on the branch path. the root node r will send PSC packet to its peer leaf node L1 or L2 immediately, then switch into its p2p protection path to transport the traffic. while the leaf node L1 or L2 will select the protection path to receive the traffic. but for the protection mechanism, the more the leaf nodes are, the more its p2p protection paths are pre-configured.

1.5. 1:n p2mp shared protection

The protection will pre-configure one p2mp shared protection path to protect multiple p2mp working paths which maybe have different end points. so the leaf nodes of the p2mp shared protection path are all leaf nodes of these protected p2mp working paths. when a defect is

detected on a working path, the root node of the protection path will send the existing extensive PSC protocol packet to its leaf nodes to identify which p2mp working path will be selected to be protected. when all the leaf nodes receive the PSC packet and decide how to process the traffic packet from the p2mp shared protection path. just as figure 3.



NOTE :

```
@@@@@: p2mp shared protection path
```

```
+++++: p2mp working path 1
```

```

#####: p2mp working path 2

```

Figure 3

a single p2mp shared protection path is used to protect the p2mp working path 1 and the p2mp working path 2. node L1 and L2 are the

sink leaf nodes of the p2mp working path 1. while node L2 and L3 are the sink leaf nodes of the p2mp working path 2. so the leaf nodes of the p2mp shared protection path are node L1 , L2 and L3. when a defect is detected on both the working path 1 and the working path 2 at the same time. The root node r of the p2mp shared protection path will select the higher priority working path to be protected. then it sends extensive PSC protocol packet to all leaf nodes L1, L2 and L3 by the p2mp shared protection path. Then the leaf nodes L1, L2 and L3 can judge whether to receive the traffic from the protection path based on the extensive PSC protocol packet.

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 State Coordination

SD: Signal Degrade

SF: Signal Fail

NR No Request

MPLS: Multi-Protocol Label Switching

MPLS-TP: Multi-Protocol Label Switching Transport Profile

3. Coordination protocol

Some protection mechanisms in this document need to PSC protocol to coordinate the switch state between the end-points of a protection domain. in order to gain to consistent solution for this coordination between the end-points of the protection domain. it will apply the existing PSC protocol to be defined in MPLS-TP Linear

protection(RFC6378) in the 1:1 p2mp protection scenario including 1:1 p2mp per-tree protection , 1:1 P2MP branch path protection. The PSC protocol in detail maybe reference to the MPLS-TP linear protection(RFC6378). while 1:n shared p2mp protection may use the extensive PSC protocol defined in the MPLS-TP 1:N protection(draft-ietf-mpls-tp-lton-protection) to implement the protection.so the procedure of extensive PSC protocol maybe reference to the mpls-tp 1:N protection document. while the 1:n shared p2mp protection only use single-phase protocol which is different from the 1:n protection for p2p path. so it is only non-locking operation and the value of L field can't be set. in addition, it is unnecessary for the protection to wait for peer node's Acknowledge(WFA), so the parameter of WFA timer isn't necessary in this document.

4. switch operation

In all of the above protection mechanism, Firstly it must pre-configure protection path to protect one or multiple p2mp working path. then it detects whether to have a defect on the working path. if a defect is detected, then the end point of the p2mp working path will switch the traffic from the working path to the protection path

4.1. 1+1 protection operation

The procedure of 1+1 protection operation is the following in detail:

1 Under normal condition, the root node bridges the p2mp traffic into the working path and the protection path, while its leaf nodes will select the p2mp working path to receive the traffic.

2 a defect is detected on the working path, some leaf nodes will select the protection path to receive the traffic which has defect on the branch path.

4.2. 1:1 p2mp per-tree protection operation

The procedure of this protection switch is the following phase:

1 under normal condition, the root node and leaf nodes will send and receive the traffic from the working path.the root node will send NR(0,0) to all leaf nodes by the protection path.

2 when a defect is already detected on the working path, the root node will bridge from the working path to the protection path. then send SF(1,1) to all leaf nodes by the protection path.

3 The leaf nodes receive the SF(1,1), All of them will switch into

the protection path to receive the traffic.

4.3. 1:1 p2mp per-leaf protection operation

The procedure of the protection is the following in detail:

1 under normal condition, the root node will select the working path to send the traffic, and the leaf node select the working path to receive the traffic.

2 when a defect is already detected on the branch path of the p2mp working path. the root node will bridge both the working path and the protection path. the sink node will select one of the two paths to receive the traffic based on whether a defect happens on the working path.

4.4. 1:1 p2mp branch path protection operation

The procedure of the protection operation is the following in detail:

1 under normal condition, the root node and all leaf nodes will select the p2mp working path to send and receive the traffic. and the root node sends NR(0,0) to each leaf node by its p2p protection path.

2 when a defect is detected on a branch path of the p2mp working path, The root node will bridge the traffic into its corresponding p2p protection path and send SF(1,1) to the leaf node of the branch path.

3 The leaf node receives the SF(1,1) packet, then it will switch from the working path to the p2p protection path to receive the traffic. other leaf nodes will still receive the traffic from the working path.

4.5. 1:n p2mp shared protection operation

the procedure of this protection operation is the following :

1 Under normal condition, the root node and all leaf nodes will select the p2mp working path to send and receive the traffic, the root node sends NR(0,0) to all leaf nodes of the p2mp protection path periodically.

2 a defect is detected on one or multiple p2mp working path, the root node will select the highest priority working path to be protected. and send SF(n,n) to all the leaf nodes of the protection path. here n indicates the index of the p2mp working path to be protected. at the same time, the root node begin to send the traffic

to be selected on the p2mp shared protection path.

3 when all leaf nodes of the p2mp shared protection path receive the SF(n,n) packet. they will know whether or not to receive the traffic from the p2mp shared protection path based on SF(n,n) packet. if a node is a leaf node of the p2mp working path to be protected, it will receive the traffic , or not, it directly abandon the traffic from the p2mp shared protection path.

5. Security Considerations

TBD

6. IANA Considerations

TBD

7. Acknowledgments

TBD

8. References

8.1. Normative References

- [RFC 5654]
IETF, "IETF RFC5654(MPLS-TP requirement)", September 2009.
- [RFC 5921]
IETF, "IETF RFC5654(MPLS-TP framework)", July 2010.
- [RFC 6372]
IETF, "MPLS Transport Profile (MPLS-TP) Survivability Framework", September 2011.
- [RFC 6378]
IETF, "MPLS Transport Profile (MPLS-TP) Linear Protection", November 2011.

8.2. Informative References

- [draft-ietf-mpls-tp-lton-protection-00]
E. Osborne, F. Zhang, Y. Weingarten, "MPLS-TP lton Protection", August 2012.

8.3. URL References

[MPLS-TP-22]

IETF - ITU-T Joint Working Team, "", 2008,
<<http://www.example.com/dominator.html>>.

Author's Address

Guoman Liu
ZTE Corporation
No.50, Ruanjian Road, Yuhuatai District
Nanjing 210012
P.R.China

Phone: +86 025 88014227
Email: liu.guoman@zte.com.cn

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: April 25, 2013

J. Jeganathan
H. Gredler
Y. Shen
Juniper Networks
Oct 22, 2012

RSVP-TE LSP egress fast-protection
draft-minto-rsvp-lsp-egress-fast-protection-01

Abstract

RFC4090 defines a fast reroute mechanism for locally repairing an RSVP-TE LSP in the order of 10s of milliseconds, in the event of a downstream link or node failure. However, this mechanism does not provide node protection for LSP egress nodes, even when an alternate egress node (a backup egress) is available that could carry the traffic to its ultimate destination. This document addresses this scenario and describes how to provide egress protection by establishing a bypass LSP from the penultimate-hop node of a LSP to the backup egress node. The methods described in this document enable local repair in the order of 10s of milliseconds, in the event of the egress node failure. These methods are only applicable if traffic carried by the LSP can be rerouted to its ultimate destination by the backup egress node.

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, 2013.

Copyright Notice

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

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

Table of Contents

| | |
|---|----|
| 1. Introduction | 3 |
| 2. Specification of Requirements | 5 |
| 3. Terminology | 5 |
| 4. Proxy method | 6 |
| 4.1. Tunnel destination Advertisement in IGP | 6 |
| 4.1.1. IS-IS proxy-node (Non-Normative) | 7 |
| 4.1.2. OSPF proxy-node (Non-Normative) | 7 |
| 4.2. Ingress Node Behavior | 7 |
| 4.3. Primary Egress Node Behavior | 8 |
| 4.4. Penultimate Hop Node | 8 |
| 4.4.1. Backup LSP Signaling during Local Repair | 8 |
| 4.5. Backup Egress Node Behavior | 8 |
| 4.5.1. Backup LSP Signaling during Local Repair | 8 |
| 4.6. Proxy method solution characteristics | 8 |
| 5. Alias model | 9 |
| 5.1. Ingress Behavior | 9 |
| 5.2. Primary Egress node | 10 |
| 5.3. Backup egress node | 10 |
| 5.3.1. Procedures for the Backup egress during Local Repair | 10 |
| 5.3.2. Processing Backup Tunnel's ERO | 10 |
| 5.4. Penultimate hop node | 10 |
| 5.4.1. Signaling a Backup Path | 10 |
| 5.4.2. Procedures for Backup Path Computation | 11 |
| 5.4.3. Signaling for Facility Protection | 11 |
| 5.4.3.1. Discovering Downstream Labels | 11 |
| 5.4.3.2. Processing Backup Tunnel's ERO | 11 |
| 5.4.3.3. PLR Procedures during Local Repair | 11 |
| 5.5. Alias method solution characterization | 11 |
| 6. Security Considerations | 11 |
| 7. Acknowledgements | 12 |
| 8. References | 12 |
| 8.1. Normative References | 12 |
| 8.2. Informative References | 13 |
| Authors' Addresses | 13 |

1. Introduction

This document describes procedures for providing fast protection for RSVP-TE LSPs in case of the egress node failure. Such protection can only be provided when an alternate egress node exists that can carry the traffic destined for the protected egress to its ultimate destination. The primary egress node of an LSP (the protected egress) terminates the LSP in steady state, while the alternate egress node (the backup egress) does so when the primary fails. A bypass LSP is established from the penultimate-hop node to the backup egress. The penultimate-hop node, serving as a PLR (point of local repair), redirects traffic to the backup egress node of the LSP using this bypass LSP in the event of primary egress node failure.

The backup egress node forwards the traffic to its ultimate destination using methods that are beyond the scope this document. For example, backup egress node could use the service specific mechanism specified in [pwe3-endpoint-fast-protection] and [l3vpn-egress PE-fast-protection] and mirror the inner labels (e.g. layer-2/3 VPN service labels) from the primary on the backup. The backup would then repair the traffic to its destination based on the mirrored labels. This document focuses on the methods for setting up the bypass LSP to the backup egress so that service specific mechanism could build top on this.

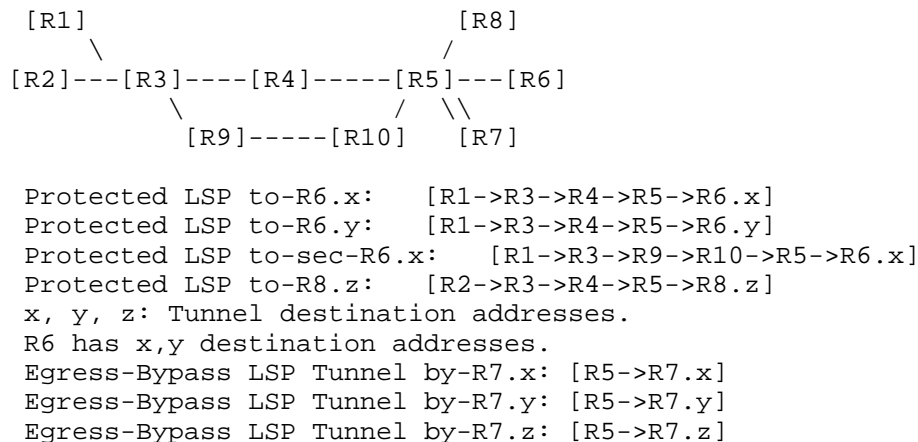


Figure 1

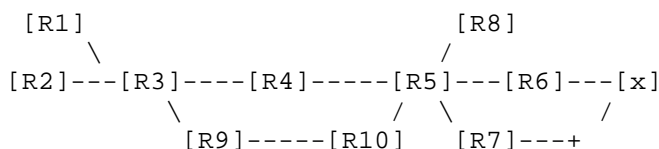
In Figure 1, four LSPs require egress protection. R6 and R8 are the primary egresses. R7 is backup egress for both R6 and R8. R5 is the penultimate hop node. R5 establishes a bypass LSP to R7 to provide fast protection in case R6 or R8 fail. Table 1 shows the bypass LSPs for each of the protected LSPs at R5.

| Protected LSP | Egress Bypass LSP |
|---------------|-------------------|
| to-R6.x | by-R7.x |
| to-R6.y | by-R7.y |
| to-sec-R6.x | by-R7.x |
| to-R8.z | by-R7.z |

Table 1

This draft describes two methods for setting up the bypass LSP to the backup egress node, the proxy node method and the alias method.

In the proxy method, an LSP endpoint address is represented as a virtual node in the TE domain, attached to the primary egress node and the backup egress node via bidirectional point-to-point TE links.



x: Tunnel destination addresses in the proxy method.

Figure 2

With the proxy method, when providing egress protection to the LSPs with destination address x, terminating on primary R6, with backup egress R7, from Figure 1, the topology is modeled as shown in Figure 2.

With this representation, penultimate-hop node R5 could use RFC 4090 RSVP fast-reroute PLR procedures to set up a bypass LSP to the backup egress node R7, by avoiding the primary egress node R6.

In alias method, an LSP endpoint address is associated with a primary egress and a explicit backup egress. The penultimate-hop node of the protected LSP may learn the backup for the LSP from backup egress IGP advertisement or by a local configuration. With this method, the penultimate-hop node can set up a bypass LSP to the backup egress node, by avoiding the primary egress node.

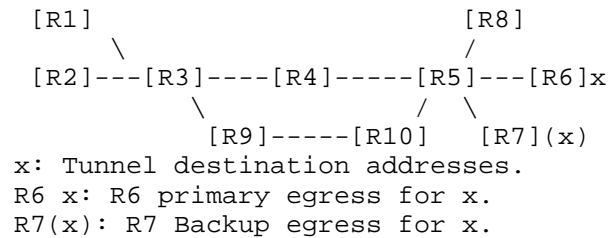


Figure 3

In Figure 3, let say x is tunnel destination address and R6 advertise x as secondary loopback address. With this alias representation R5 see the x as x{R6,R7} where R6 is primary and R7 is backup for x. This primary to backup mapping is either learn through R7's IGP backup availability advertisement or by a local configuration in R5.

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

PLR: Point of Local Repair. The head-end LSR of a backup tunnel or a detour LSP

PHN: Penultimate Hop Node for an LSP.

Primary egress node: Node terminates a LSP in steady state.

Primary: Primary egress node.

Egress Protected LSP: A Protected LSP that also required protection from primary egress node failure

Backup egress node: Node could rerouted/repaired data carried in a protected LSP

Backup node: Backup egress node.

Protector: Backup egress node.

Protector and Backup node are used interchangeably but convey the same meaning.

4. Proxy method

In this method, an LSP endpoint address is represented as a virtual TE node connected to a primary egress node and a backup egress node with bidirectional TE links, as shown in Figure 2. With this model, node protection establishment and bypass LSP path computation on the penultimate hop of an LSP can follow the procedure described in RFC4090.

4.1. Tunnel destination Advertisement in IGP

Advertising the tunnel destination as a stub proxy TE node requires two steps: 1) a node representation (proxy-node) and 2) links to and from primary egress and backup egress.

The primary advertises a proxy node with two links, to the primary egress and the backup egress, respectively. The router ID of the proxy node is LSP end point address. The system-ID of the proxy is derived from the LSP end point address with BCD encoding. The resulting system-ID and router-ID MUST be unique within the IGP routing domain.

Both stub links are advertised with maximum routable metric and TE metric, and zero bandwidth, as shown in Figure 4. This avoids the proxy node serving as a transit node for any path. The router-ID or system-ID of the backup egress can be dynamically learned from the IGP link state database or can be configured on the primary egress.

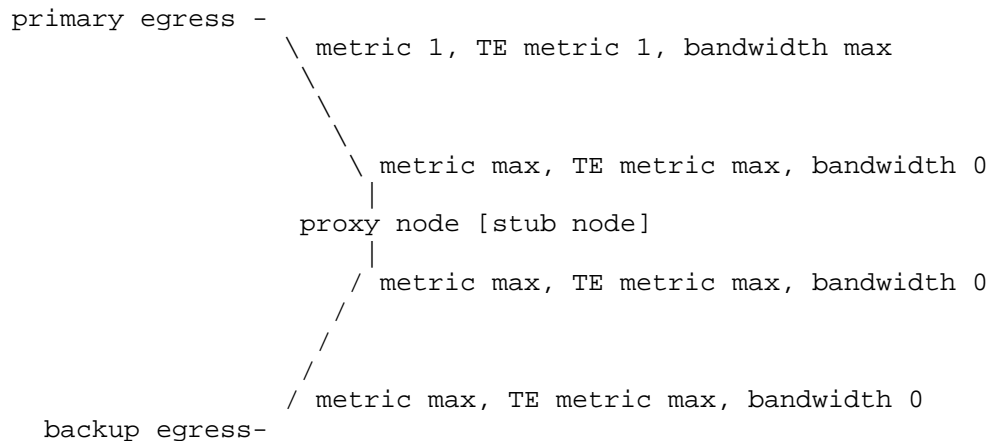


Figure 4

The primary egress advertises an unnumbered transit link to the proxy

node, with metric 1, TE metric 1, and maximum bandwidth. It may be necessary for the primary node to have the capabilities to advertise multiple TE unnumbered transit links between primary node and proxy-node. The upper bound on the number of such links is the number of the links the primary node advertises into TE.

The backup egress advertises an unnumbered transit link to the proxy node, with MAX metric, MAX TE metric, and zero bandwidth. Other TE characteristic of the links can be configured and advertised as well.

4.1.1. IS-IS proxy-node (Non-Normative)

When IS-IS is used as IGP to advertise the proxy node, only zeroth fragment of the proxy-node advertisement is valid. All other fragments SHOULD be ignored. The zeroth fragment MUST include the area address TLV and MAY include the hostname TLV.

The set of area addresses advertised in proxy node zeroth fragment link-state PDU MUST be a subset of Area Addresses advertised by the primary egress in the zeroth fragment of the link-state PDU of the corresponding IS-IS level. The advertisement SHOULD be syntactically identical to the primary egress zeroth fragment at corresponding IS-IS level. The hostname SHOULD be set as <tunnel-destination + primary egress hostname>.

The Overload (OL) MUST be set to 1. The Attached (ATT), and Partition Repair (P) bits MUST be set to 0.

4.1.2. OSPF proxy-node (Non-Normative)

The advertising router and Link State ID of router LSA MUST be LSP end point address. All options bits in router LSA MUST be set to zero.

4.2. Ingress Node Behavior

The ingress node of an LSP requesting egress protection SHOULD follow the procedures described in RFC 2205 and RFC 4090 to signal the LSP. In particular, it SHOULD set the destination to the endpoint address (i.e. the proxy node), and the "link protection desired" flag and the "node protection desired" flag in the SESSION_ATTRIBUTE object of the Path message. In path computation, it MAY optionally exclude MAX metric links to avoid the link between the backup egress and the proxy node.

4.3. Primary Egress Node Behavior

When the primary egress node receives a Path message for the LSP with destination matching the proxy node address, it MUST append two entities in the RRO object of Resv message: 1) the proxy node as a virtual downstream node, and 2) itself as a virtual transit node. The entity for the proxy node is encoded as {proxy node address, proxy link ID, implicit NULL}.

4.4. Penultimate Hop Node

When the penultimate hop node receives a Resv message from the primary egress, it sees itself as two hops away from LSP's destination rather than one hop, based on the RRO. Thus, it can set up node protection for the LSP by following the procedure described in RFC 4090. It SHOULD set up a bypass LSP to the backup egress node. When computing a path for the bypass LSP, it SHOULD avoid the primary egress node and choose a path via the backup egress node to reach the proxy node.

4.4.1. Backup LSP Signaling during Local Repair

The penultimate hop node SHOULD use the same procedure as defined RFC4090 to signal the backup Path, in the event of failure of the primary egress node.

4.5. Backup Egress Node Behavior

When the backup egress node receives Path message of the bypass LSP, it MUST terminate the Path message based on match between the LSP destination and the proxy node address. It SHOULD assign a non-reserved label to the bypass LSP. This non-reserved label provide forwarding context during repair.

4.5.1. Backup LSP Signaling during Local Repair

During local repair, the backup egress node will receive Path message of egress-protected LSP from the penultimate hop node. The backup egress node SHOULD terminate the Path message, and respond with a Resv message.

4.6. Proxy method solution characteristics

The biggest advantage of the proxy method is that it does not require protocol extensions and can be implemented locally at the tunnel egress node. Thus, no software upgrades are required in the core of the network.

The proxy method has the following caveats:

1. To support TE constrains like colors and SRLG for a protected LSP the primary needs to have the capability to advertise multiple links to between proxy and primary.
2. Bypass LSP with constrains cannot be supported.
3. If IS-IS is used as the IGP then the Primary node should not be configured with overload bit.
4. Backup egress could be used as primary end point in the forwarding plane if the protected LSP established to backup instead of primary in transient condition.

Due to its characteristics, the proxy method is suitable for mixed environments, where an upgrade of the entire network is not feasible.

5. Alias model

In this model Penultimate hop node (PHN) of a protected LSP understands that LSP end point has a backup egress and it could repair traffic carried in the protected LSP in the event of primary egress failure. After the primary egress failure, the PHN reroutes traffic using a bypass tunnel to backup egress. The tunnel endpoint address and backup egress mapping could be configured in penultimate hop node or signaled through IGP from the backup. Table 2 illustrates the PHN mapping primary to backup mapping for the topology in Figure 1.

| Primary Egress
Router ID | Backup egress
router ID | Backup LSP destination
address. |
|-----------------------------|----------------------------|------------------------------------|
| 10.1.2.6 | 10.1.1.6 | 10.1.1.7 |
| 10.1.2.6 | 10.1.3.6 | 10.1.1.6 |
| 10.1.1.7 | 10.1.3.6 | 10.1.2.8 |
| 10.1.1.8 | 10.1.1.7 | 10.1.2.8 |

Table 2: Table mapping

5.1. Ingress Behavior

The ingress should follow the procedure in RFC 3209 with tunnel endpoint address. The path computation could use the tunnel endpoint address advertised using the procedures of RFC 5786.

5.2. Primary Egress node

Primary egress node advertises tunnel end points that require protection using RFC 5786 in OSPF and/or IP interface addresses TLV(132) in IS-IS. These TLVs are defined as Local address advertisement in TE. The rest of the behavior is same RFC 4090.

5.3. Backup egress node

When backup receives a Path message not through a bypass tunnel for a destination address it protects with ERO constrains only one self sub objects then it MUST accept and respond with RRO objects in Resv message. The RRO object {node ID, Ip address, label} for tunnel end address set with {Node ID, tunnel endpoint address, non-reserved label}. This non-reserved label provide forwarding context during local repair.

5.3.1. Procedures for the Backup egress during Local Repair

The Backup egress sends Resv, ResvTear, and PathErr messages by sending them directly to the address in the RSVP_HOP object, as specified in [RSVP-TE].

5.3.2. Processing Backup Tunnel's ERO

When backup receive Path message through a bypass tunnel with one sub-object for destination address it protects then it should accept ERO.

5.4. Penultimate hop node

PLR learns/configured backup egress for tunnel a end point address advertised by primary egress. When PLR setup bypass for node protection LSP it will also lookup for the backup egress if PLR is penultimate hop of the LSP. If backup egress is available for LSP tunnel end point address then it setup bypass-LSP to backup egress if it is not setup already. The constrains will be exclude egress node. PHN could setup bypass-LSP with destination as backup egress node or tunnel endpoint address. If the bypass tunnel endpoint address is not the protected LSP tunnel endpoint then it also initiates backup LSP for tunnel end point address through bypass tunnel to learn the label to use in failure.

5.4.1. Signaling a Backup Path

PHP SHALL use the same procedure as defined RFC4090 to signal the backup Path.

5.4.2. Procedures for Backup Path Computation

PLR has to find the desired explicit route for the backup path. This can be done using a CSPF computation. If PLR is PHN for the protected LSP needs node protection then destination for the backup path MUST be backup egress router ID with the constraint that the LSP cannot traverse the primary egress node and/or link whose failure is being protected against. For other constraints SHOULD follow RFC4090.

5.4.3. Signaling for Facility Protection

A PHN use one or more bypass tunnels to protect against the failure of a egress primary node. This bypass tunnels set up in advance or dynamically created as new protected LSPs are signaled.

5.4.3.1. Discovering Downstream Labels

To support facility backup, the PHN must determine the label that will indicate to the backup egress that packets received with that label should be processed by primary egress context. This can be done by setting up the UHP bypass tunnel to the backup egress with tunnel endpoint address as destination.

5.4.3.2. Processing Backup Tunnel's ERO

Sub-objects belonging to abstract nodes that precede the tunnel endpoint Point are removed. A sub-object identifying the Backup Tunnel destination is then added.

5.4.3.3. PLR Procedures during Local Repair

PHN SHALL use the procedures defined in RFC4090 during the local repair.

5.5. Alias method solution characterization

The alias method will work with arbitrary TE constraints and suitable for networks that required LSP with those TE constraints. To avoid PLR static backup egress configuration, IGP extension is required.

6. Security Considerations

The security considerations discussed in RFC 5036, RFC 5331, RFC 3209, and RFC 4090 apply to this document.

7. Acknowledgements

This document leverages work done by Hannes Gredler, Yakov Rekhter and several others on LSP tail-end protection. Thanks to Ina Minei, Nischal Sheth, Nitin Bahadur, Ashwin Sampath and Kaliraj Vairavakkalai for their contribution.

8. References

8.1. Normative References

- [RFC5331] Aggarwal, R., Rekhter, Y., and E. Rosen, "MPLS Upstream Label Assignment and Context-Specific Label Space", RFC 5331, August 2008.
- [RFC4364] Rosen, E. and Y. Rekhter, "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.

8.2. Informative References

- [RFC5286] Atlas, A. and A. Zinin, "Basic Specification for IP Fast Reroute: Loop-Free Alternates", RFC 5286, September 2008.
- [RFC5714] Shand, M. and S. Bryant, "IP Fast Reroute Framework", RFC 5714, January 2010.
- [pwe3-endpoint-fast-protection] Shen, Y., Ed. and Aggarwal, R., "PW Endpoint Fast Failure Protection", 2011, <pwe3-endpoint-fast-protection>.
- [l3vpn-egress-PE-fast-protection] Jeganathan, J. and G. Gredler, "2547 egress PE Fast Failure Protection", 2011, <2547-egress-PE-fast-protection>.

Authors' Addresses

Jeyanthan Minto Jeganathan
Juniper Networks
1194 N Mathilda Avenue
Sunnyvale, CA 94089
USA

Email: minto@juniper.net

Hannes Gredler
Juniper Networks
1194 N Mathilda Avenue
Sunnyvale, CA 94089
USA

Email: hannes@juniper.net

Yimin Shen
Juniper Networks
10 Technology Park Drive
Westford, MA 01886
USA

Email: yshen@juniper.net

Internet Engineering Task Force
Internet-Draft
Intended status: Standards Track
Expires: March 10, 2013

Y. Shen
Juniper Networks
Y. Kamite
NTT Communications Corporation
September 6, 2012

RSVP Setup Protection
draft-shen-mpls-rsvp-setup-protection-01

Abstract

RFC 4090 specifies an RSVP facility-backup fast reroute mechanism that can protect LSPs against link and node failures. This document extends the mechanism to provide "setup protection" for LSPs during initial Path message signaling time. In particular, it enables a router to reroute an LSP via an existing bypass LSP, when there is a link or node failure along the desired path.

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 10, 2013.

Copyright Notice

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

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

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

Table of Contents

| | |
|--|----|
| 1. Introduction | 3 |
| 2. Specification of Requirements | 4 |
| 3. Theory of Operation | 4 |
| 3.1. New RSVP Attribute Flag | 5 |
| 3.2. New RSVP Attributes TLVs | 5 |
| 3.2.1. Protected LSP Sender IPv4 Address TLV | 6 |
| 3.2.2. Protected LSP Sender IPv6 Address TLV | 6 |
| 3.3. PLR behavior | 7 |
| 3.4. MP behavior | 9 |
| 3.5. Local Revertive Mode | 9 |
| 4. IANA Considerations | 10 |
| 5. Security Considerations | 10 |
| 6. Acknowledgements | 10 |
| 7. References | 10 |
| 7.1. Normative References | 10 |
| 7.2. Informative References | 11 |
| Authors' Addresses | 11 |

1. Introduction

In RSVP facility-backup fast reroute (FRR) [RFC 4090], the router at a point of local repair (PLR) of an LSP can redirect traffic via a bypass LSP upon a failure of the immediate downstream link or node. Such kind of protection is normally established after the PLR has received a Resv message of the LSP. In link protection, the PLR must learn the label and address of the next-hop router, before it can set up or select a bypass LSP to protect the LSP. Likewise, in node protection, the PLR must learn the label and address of the next-next-hop router. The information of the label and the address is carried in the Resv message.

Imagine a scenario where an LSP is being signaled, but its Path message carries an EXPLICIT_ROUTE object (ERO) that is statically configured or computed based on a topology that may not reflect the current state of every link or node of the network. If a link or node along this path is in a failure condition, RSVP signaling will halt at the router immediate upstream of the failure. This will be the case even if there is an existing bypass LSP protecting the link or node for some other LSPs. In other words, the LSP is not protected during its setup time, i.e. the initial Path message signaling time.

In this situation, the network would rely on IGP to flood the up-to-date traffic engineering (TE) information, and the router immediate upstream of the failure to send a PathErr message to notify the ingress router. The ingress router can then compute and signal a new path to avoid the failed link or node. However, this approach may not always be possible or desirable, as in the scenarios below.

1. Pre-configured or pre-defined paths. If the path is pre-configured or pre-defined, and the ingress router is incapable of computing a new path, the LSP will not be set up.
2. LSPs with a strict requirement for setup time. IGP TE information flooding, PathErr message propagation, path re-computation, and RSVP re-signaling may introduce a significant delay to LSP establishment. This may impact on signaling performance for services that have a strict requirement for LSP setup time, such as an on-demand transport service for real-time data.
3. Sibling P2MP sub-LSPs sharing a failed link. In this case, the LSP being signaled is a sub-LSP of a P2MP LSP, and it is supposed to share the failed link with an existing sibling sub-LSP (i.e. another sub-LSP of the same P2MP LSP) which is being protected by a bypass LSP. If the new sub-LSP is rerouted via a different

path, it will not be able to share the data flow over the bypass LSP with that sibling sub-LSP, and unnecessary traffic flow will be generated in the network.

This document extends the RSVP facility-backup fast reroute mechanism to provide so-called "setup protection" for LSPs. During the initial Path message signaling of an LSP, if there is a link or node failure along the desired path, and if there is a bypass LSP protecting the link or node, the LSP will be signaled through the bypass LSP. The LSP will be established as if it was originally set up along the desired primary path and then failed over to the bypass LSP after the link or node failure. When the link or node is restored, the LSP MAY be reverted to the primary path. The mechanism supports both P2P and P2MP LSPs.

2. Specification of Requirements

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

3. Theory of Operation

When an LSP is being signaled by RSVP, a Path message is sent hop by hop from the ingress router to the egress router, following the path defined by an ERO. The setup protection mechanism in this document enables an ingress or transit router to reroute the LSP via a bypass LSP, if the router detects a failure of the immediate downstream link or node represented by the next hop in the ERO, i.e. next ERO hop. This router is referred to as a PLR.

The mechanism is relevant when the Path message carries the "local protection desired" flag in the SESSION_ATTRIBUTE object [RFC 4090] and a new "setup protection desired" flag defined in this document (Section 3.1).

On a PLR, the mechanism is only applicable when the next ERO hop is a strict hop, and in case of node protection, the next-next ERO hop is also a strict hop. A strict next ERO hop allows the PLR to unambiguously decide the intended downstream link or node on the desired path, and hence reliably detect the status of the link or node. In link protection, the strict next ERO hop also indicates the merge point (MP), i.e. the destination of the bypass LSP to be used for rerouting the LSP. In node protection, the strict next-next ERO hop indicates the MP.

When performing setup protection, the PLR signals a backup LSP by tunneling a Path message through the bypass LSP. Like the Path message of a backup LSP in the normal facility-backup FRR, this Path message carries an address of the PLR as the sender address. In addition, the Path message also carries some information of the protected LSP (Section 3.2). When the MP receives the Path message, it terminates the backup LSP, and then re-creates the protected LSP. If the MP is a transit router of the protected LSP, it signals the LSP further downstream.

Eventually, the LSP will be established end to end, with the backup LSP tunneled through the bypass LSP from the PLR to the MP. The RSVP state on the PLR and the MP and the RSVP messages generated by these routers are no different than those of an LSP in a post-failure situation of the normal facility-backup FRR.

After the link or node is restored, the PLR MAY revert the LSP to the primary path, in the same manner as the local revertive mode specified in [RFC 4090].

The setup protection mode MAY be enabled and disabled on a router based on configuration. For an LSP to be setup-protected, the mode MUST be enabled on both PLR and MP. If it is enabled on a PLR but disabled on an MP, the MP SHOULD reject the Path message of the backup LSP and send a PathErr message, as described Section 3.4.

3.1. New RSVP Attribute Flag

In order to facilitate explicit request for setup protection, this document defines a new "setup protection desired" flag in the Attribute Flags TLV, which is carried in the LSP_ATTRIBUTES object [RFC5420] of the Path message of a protected LSP.

3.2. New RSVP Attributes TLVs

This document defines two new RSVP Attributes TLVs [RFC 5420]. They are used by a PLR to convey to an MP the original sender address of a protected LSP. Both TLVs are carried in the LSP_REQUIRED_ATTRIBUTES object in the Path message of a backup LSP.

- o Protected LSP Sender IPv4 Address TLV
- o Protected LSP Sender IPv6 Address TLV

3.2.1. Protected LSP Sender IPv4 Address TLV

The Protected LSP Sender IPv4 Address TLV is defined with type X. It is allowed on LSP_REQUIRED_ATTRIBUTES object, and not allowed on LSP_ATTRIBUTES object. It is encoded as the following.

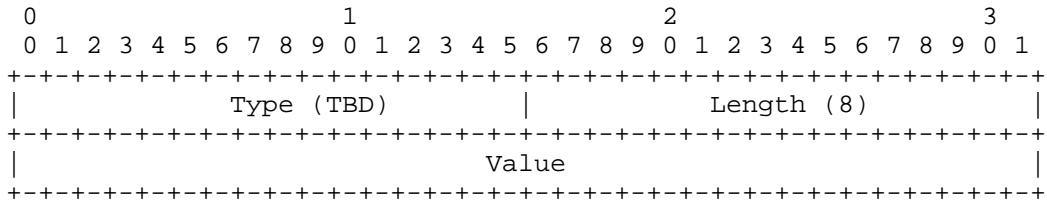


Figure 1

Type

TBD

Length

8

Value

Original sender address in the IPv4 SENDER_TEMPLATE object of the protected LSP.

3.2.2. Protected LSP Sender IPv6 Address TLV

The Protected LSP Sender IPv6 Address TLV is defined with type Y. It is allowed on LSP_REQUIRED_ATTRIBUTES object, and not allowed on LSP_ATTRIBUTES object. It is encoded as the following.

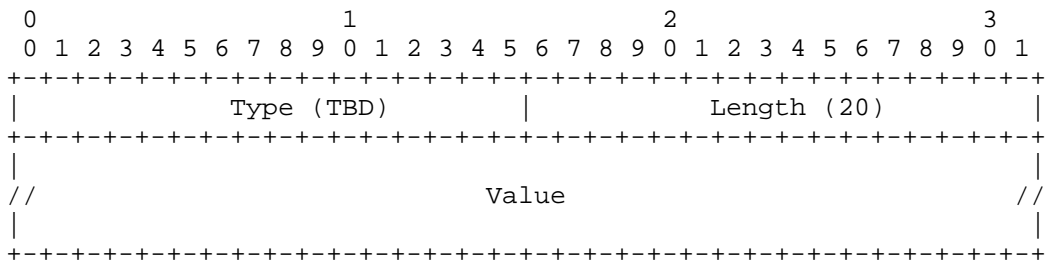


Figure 2

Type

TBD

Length

20

Value

Original sender address in the IPv6 SENDER_TEMPLATE object of the protected LSP.

3.3. PLR behavior

When a router has a Path message to send out, if the Path message carries the "local protection desired" flag in the SESSION_ATTRIBUTE object and the "setup protection desired" flag in the LSP_ATTRIBUTES object, and if the next ERO hop is a strict IPv4 or IPv6 prefix, the router SHOULD validate the prefix against the routing table, the traffic engineering (TE) database, and/or a topology database. If the prefix is reachable and is one hop away from the router, the Path message is sent as it is. Otherwise, there is a possibility that the link or node represented by the prefix has experienced a failure.

The router SHOULD determine this by searching for an existing bypass LSP that is protecting the prefix. If the protected LSP desires link protection, the destination of the bypass LSP (i.e. MP) is considered as the router that owns the prefix. If the LSP desires node protection with the "node protection desired" flag set in the SESSION_ATTRIBUTE object, the next-next ERO hop of the LSP must also be a strict prefix, and the MP is considered as the router that owns this prefix.

If a bypass LSP is not found, the router MUST originate a PathErr with code = 24 (routing problem) and sub-code = 2 (bad strict node).

If a bypass LSP is found, the router MUST act as a PLR of setup protection, and reroute the protected LSP via the bypass LSP. If multiple satisfactory bypass LSPs exist, the PLR MAY select one based on bandwidth constraints or local policies. If the protected LSP is a sub-LSP of a P2MP LSP, a bypass LSP that is protecting an existing sibling sub-LSP MUST be preferred, in order to minimize traffic duplication in the network.

The PLR SHOULD NOT send a Path message for the protected LSP. Instead, it MUST create a backup LSP, and send a Path message of the backup LSP to the MP via the bypass LSP. The Path message is constructed by using the sender template specific method [RFC 4090]. In particular, it has the sender address in the SENDER_TEMPLATE object set to an address of the PLR. It MUST also carry an LSP_REQUIRED_ATTRIBUTES object containing a Protected LSP Sender IPv4 Address TLV or Protected LSP Sender IPv6 Address TLV.

Upon receiving a Resv message of the backup LSP from the MP, the PLR SHOULD bring up both of the backup LSP and the protected LSP. If the PLR is the ingress router of the protected LSP, the LSP has been set up successfully. If the PLR is a transit router, it MUST send a Resv message upstream for the protected LSP, with the "local protection available", "local protection in use", and optionally "node protection" and "bandwidth protection" flags set to 1 in the RRO hop corresponding to the PLR [RFC 4090]. The PLR SHOULD originate a PathErr message with code = 25 (notify error) and sub-code = 3 (tunnel locally repaired).

The PLR SHOULD also install a forwarding entry for the protected LSP. The next-hop of this entry MAY indicate zero, one, or two outgoing labels, depending on whether any of the backup LSP's label and the bypass LSP's label is Implicit NULL. In the case of two labels, the inner label is the backup LSP's label, and the outer label is the bypass LSP's label.

If the PLR receives a PathErr message when signaling the backup LSP, the PLR MUST NOT bring up the backup LSP or the protected LSP. If the PLR is a transit router of the protected LSP, it MUST send a PathErr message upstream for the protected LSP. Likewise, if the PLR receives a PathErr message after the backup LSP and the primary LSP have been set up, and the PLR is a transit router of the protected LSP, it MUST also send a PathErr message upstream for the protected LSP.

When the PLR receives a ResvTear message of the backup LSP, the PLR MUST bring down both the backup LSP and the protected LSP. If the PLR is a transit router of the protected LSP, it MUST send a ResvTear message upstream for the protected LSP.

In any cases where the PLR tears down the protected LSP due to a received PathTear message, RSVP state time-out, configuration change, administrative command, etc, the PLR MUST also tear down the backup LSP by sending a PathTear message through the bypass LSP.

3.4. MP behavior

When an MP receives the Path message of a backup LSP, it SHOULD detect the setup protection condition based on the presence of Protected LSP Sender IPv4 Address TLV or Protected LSP Sender IPv6 Address TLV in LSP_REQUIRED_ATTRIBUTES object.

If setup protection mode is disabled on the MP, it MUST reject the Path message, by sending a PathErr with code = 2 (policy control failure) to the PLR.

Otherwise, the MP MUST terminate the backup LSP and re-create the protected LSP. If the MP is the egress router of the protected LSP, it MUST also terminate the protected LSP. If the MP is a transit router of the LSP, it MUST send a Path message downstream for the protected LSP. The Path message has the sender address in SENDER_TEMPLATE object set to the original address of the ingress router, based on the above received TLV. The Path message MUST NOT carry the Protected LSP Sender IPv4 Address TLV or Protected LSP Sender IPv6 Address TLV received in the above LSP_REQUIRED_ATTRIBUTES object.

The MP MUST allocate a label for the backup LSP, and distribute it to the PLR via the Resv message of the backup LSP. If the protected LSP is a sub-LSP of a P2MP LSP and there is an existing sibling sub-LSP, the MP SHOULD allocate the same label as the sibling sub-LSP, in order to avoid traffic duplication in the network.

When the MP receives a PathTear message of the backup LSP, it MUST tear down both the backup LSP and the protected LSP. If the MP is a transit router of the protected LSP, it MUST send a PathTear message downstream.

In any cases where the MP receives or originates a PathErr or ResvTear message for the protected LSP, the MP SHOULD translate the message to a message of the backup LSP and send it to the PLR.

3.5. Local Revertive Mode

When the failed link or node is restored, the PLR MAY revert the protected LSP to its desired primary path, by following the procedure of local revertive mode described in [RFC 4090].

4. IANA Considerations

This document defines a new flag in the Attribute Flags TLV, which is carried in the LSP_ATTRIBUTES Object of Path message. This flag is used to communicate whether setup protection is desired for an LSP. New flag value needs to be assigned to it by IANA.

Setup Protection Desired: TBD

This document defines two new RSVP Attributes TLVs, which are carried in the LSP_REQUIRED_ATTRIBUTES object of Path message. New type values need to be assigned to them by IANA.

Protected LSP Sender IPv4 Address TLV

Protected LSP Sender IPv6 Address TLV

5. Security Considerations

The security considerations discussed in RFC 3209, RFC 4090 and RFC 4875 apply to this document.

6. Acknowledgements

Thanks to Rahul Aggarwal, Disha Chopra, and Nischal Sheth for their contribution.

7. References

7.1. Normative References

- [RFC2205] Braden, B., Zhang, L., Berson, S., Herzog, S., and S. Jamin, "Resource ReSerVation Protocol (RSVP) -- Version 1 Functional Specification", RFC 2205, September 1997.
- [RFC3209] Awduche, D., Berger, L., Gan, D., Li, T., Srinivasan, V., and G. Swallow, "RSVP-TE: Extensions to RSVP for LSP Tunnels", RFC 3209, December 2001.
- [RFC4090] Pan, P., Swallow, G., and A. Atlas, "Fast Reroute Extensions to RSVP-TE for LSP Tunnels", RFC 4090, May 2005.
- [RFC5420] Farrel, A., Papadimitriou, D., Vasseur, JP., and A. Ayyangarps, "Encoding of Attributes for MPLS LSP

Establishment Using Resource Reservation Protocol Traffic Engineering (RSVP-TE)", RFC 5420, February 2009.

- [RFC4875] Aggarwal, R., Papadimitriou, D., and S. Yasukawa, "Extensions to Resource Reservation Protocol - Traffic Engineering (RSVP-TE) for Point-to-Multipoint TE Label Switched Paths (LSPs)", RFC 4875, May 2007.
- [RFC3471] Berger, L., "Generalized Multi-Protocol Label Switching (GMPLS) Signaling Functional Description", RFC 3471, January 2003.
- [RFC3472] Ashwood-Smith, P. and L. Berger, "Generalized Multi-Protocol Label Switching (GMPLS) Signaling Constraint-based Routed Label Distribution Protocol (CR-LDP) Extensions", RFC 3472, January 2003.
- [RFC3031] Rosen, E., Viswanathan, A., and R. Callon, "Multiprotocol Label Switching Architecture", RFC 3031, January 2001.

7.2. Informative References

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

Authors' Addresses

Yimin Shen
Juniper Networks
10 Technology Park Drive
Westford, MA 01886
USA

Phone: +1 9785890722
Email: yshen@juniper.net

Yuji Kamite
NTT Communications Corporation
Granpark Tower 3-4-1 Shibaura, Minato-ku
Tokyo 108-8118
Japan

Email: y.kamite@ntt.com

MPLS
Internet-Draft
Intended status: Informational
Expires: April 11, 2013

C. Villamizar, Ed.
Outer Cape Cod Network
Consulting
K. Kompella
Contrail Systems
October 8, 2012

MPLS Forwarding Compliance and Performance Requirements
draft-villamizar-mpls-forwarding-00

Abstract

This document provides guidelines for implementors regarding MPLS forwarding and a basis for evaluations of forwarding implementations. Guidelines cover basic MPLS forwarding, forwarding when a deep MPLS label stack is encountered, MPLS UHP operations which require one or more label POP plus a PUSH, guidelines for hashing an MPLS stack and payload for multipath, and conformance and performance requirements for recent pseudowire and MPLS standards.

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 11, 2013.

Copyright Notice

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

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

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

Table of Contents

| | |
|---|----|
| 1. Introduction | 3 |
| 1.1. Apparent Misconceptions | 3 |
| 1.2. Target Audience | 4 |
| 2. Forwarding Issues | 5 |
| 2.1. Forwarding Basics | 5 |
| 2.1.1. Early Uses of Multiple Label Stack Entries | 5 |
| 2.1.2. MPLS Link Bundling | 6 |
| 2.1.3. MPLS Hierarchy | 6 |
| 2.2. Packet Rates | 6 |
| 2.3. MPLS Multipath Techniques | 7 |
| 2.3.1. Pseudowire Control Word | 8 |
| 2.3.2. Pseudowire Flow Label | 8 |
| 2.3.3. MPLS Entropy Label | 8 |
| 2.4. MPLS-TP and UHP | 9 |
| 3. Questions for Suppliers | 9 |
| 4. Forwarding Compliance and Performance Testing | 11 |
| 5. IANA Considerations | 15 |
| 6. Security Considerations | 15 |
| 7. References | 15 |
| 7.1. Normative References | 15 |
| 7.2. Informative References | 16 |
| Authors' Addresses | 17 |

1. Introduction

The document addresses concerns raised on the MPLS WG mailing list about shortcomings in implementations of MPLS forwarding.

Although this document is informational, the key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" are used. For those who wish to take the advice of this document, these keywords SHOULD be interpreted as described in RFC 2119 [RFC2119]. Similarly, the References section is split into Normative and Informative subsections. In this case references which are normative for forwarding are listed as normative. References which describe signaling only, though normative with respect to signaling, are listed as informative here, as they are informative with respect to MPLS forwarding.

1.1. Apparent Misconceptions

In early generations of forwarding silicon (which may now be behind us), there apparently were some misconceptions about MPLS. The following statements may clear up some of these misconceptions.

1. There are practical reasons to have more than one or two labels in an MPLS label stack. Under some circumstances the label stack can become quite deep. See Section 2.1.
2. The label stack must be considered to be arbitrarily deep. If a the bottom of the label stack cannot be found, but sufficient number of labels exist to forward, an LSR MUST forward the packet. An LSR MUST NOT assume the packet is malformed unless the end of packet is found before bottom of stack. See Section 2.1.
3. In networks where deep label stacks are encountered, they are not rare. Full packet rate performance is required regardless of label stack depth, except where multiple POP operations are required. See Section 2.1.
4. Research has shown that long bursts of short packets with 40 byte or 44 byte common IP payload sizes in these bursts. This is due to TCP ACK compression [ACK-compression].
 - A. A forwarding engine SHOULD, if practical, be able to sustain an arbitrarily long sequence of small packets arriving at full interface rate.

- B. If indefinite full packet rate for small packets is not practical, a forwarding engine MUST be able to buffer a long sequence of small packets inbound to the decision engine and sustain full interface rate for some reasonable average packet rate.

See Section 2.2.

- 5. For practical reasons, support for pseudowire control word SHOULD be considered mandatory by the implementor and system designer. Deployment of pseudowire control word MAY be considered optional. See Section 2.3.1.
- 6. For practical reasons, support for adding a pseudowire Flow Label [RFC6391] SHOULD be considered mandatory by the implementor and system designer. Deployment of this features MAY be considered optional. See Section 2.3.2.
- 7. For practical reasons, support for adding a MPLS Entropy Label [I-D.ietf-mpls-entropy-label] SHOULD be considered mandatory by the implementor and system designer. Deployment of this features MAY be considered optional. See Section 2.3.3.

1.2. Target Audience

This document is intended for multiple audiences: implementor (implementing MPLS forwarding in silicon or in software); systems designer (putting together a MPLS forwarding systems); deployer (running an MPLS network). These guidelines are intended to serve the following purposes:

- 1. Explain what to do and what not to do when a deep label stack is encountered. (audience: implementor)
- 2. Highlight pitfalls to look for when implementing an MPLS forwarding chip. (audience: implementor)
- 3. Provide a checklist of features and performance specifications to request. (audience: systems designer, deployer)
- 4. Provide a set of tests to perform. (audience: systems designer, deployer).

The implementor, systems designer, and deployer have a transitive supplier customer relationship. It is in the best interest of the supplier to review their product against their customer's checklist and customer's customer's checklist if applicable.

2. Forwarding Issues

A brief review of forwarding issues is provided in the subsections that follow. This section provides some background on why some of these requirements exist. The questions to ask of suppliers and testing is covered in the following sections, Section 3 and Section 4.

2.1. Forwarding Basics

Basic MPLS architecture and MPLS encapsulation, and therefore packet forwarding is defined in [RFC3031] and [RFC3032]. RFC3031 and RFC3032 are somewhat LDP centric. RSVP-TE supports traffic engineering (TE) and fast reroute, features that LDP lacks. The base document for RSVP-TE based MPLS is [RFC3209].

A few RFCs update RFC3032. Those with impact on forwarding include the following.

1. TTL processing is clarified in [RFC3443].
2. The use of MPLS Explicit NULL is modified in [RFC4182].
3. Diffserv is supported by [RFC3270] and [RFC4124]. The "EXP" field is renamed to "Traffic Class" in [RFC5462], removing any misconception that it was available for experimentation or could be ignored.
4. ECN is supported by [RFC5129].
5. The MPLS G-ACh and GAL are defined in [RFC5586].

A few RFCs update RFC3209. Those that are listed as updating RFC3209 generally impact only RSVP-TE signaling. Forwarding is modified by major extension built upon RFC3209. Some of these extensions are discussed in following subsections.

2.1.1. Early Uses of Multiple Label Stack Entries

MPLS deployments in the early part of the prior decade (circa 2000) tended to support either LDP or RSVP-TE. LDP was favored by some for its ability to scale close to the network edges without adding deployment complexity. RSVP-TE was favored where traffic engineering or fast reroute were considered important.

The use of MPLS FRR [RFC4090] added a second label to MPLS traffic, but only when FRR protection was in use.

At least one major service provider made use of LDP over RSVP-TE in their core network in the circa 2000-2005 time frame. LDP supported VPN services to the provider edges. RSVP-TE provided TE and FRR in the core. This yields two labels on nearly all packets in the core. They also used FRR which yields three labels on a large subset of traffic while FRR protection is active. VPNs added yet another label, bringing the label stack depth (with FRR) to four.

2.1.2. MPLS Link Bundling

MPLS Link Bundling was the first RFC to address the need for multiple parallel links between nodes [RFC4201]. MPLS Link Bundling is notable in that it tried not to change MPLS forwarding, except in specifying the "All-Ones" component link. MPLS Link Bundling is seldom if ever deployed. Instead multipath techniques described in Section 2.3 are used.

2.1.3. MPLS Hierarchy

MPLS hierarchy is defined in [RFC4206]. Although RFC4206 is considered part of GMPLS, the Packet Switching Capable (PSC) portion of the MPLS hierarchy are applicable to MPLS and may be supported in an otherwise GMPLS free implementation. The MPLS PSC hierarchy remains the most likely means of providing further scaling in an RSVP-TE MPLS network, particularly where the network is designed to provide RSVP-TE connectivity to the edges. This is the case for envisioned MPLS-TP networks. The use of the MPLS PSC hierarchy can add as many as four labels to a label stack, though it is likely that only one layer of PSC will be used in the near future.

2.2. Packet Rates

While average packet size of Internet traffic may be large, long sequences of small packets have both been predicted in theory and observed in practice. Traffic compression and TCP ACK compression can conspire to create long sequences of packets of 40-44 bytes in payload length. If carried over Ethernet, the 64 byte minimum payload applies, yielding a packet rate of approximately 150 Mpps (million packets per second) for the duration of the burst. The peak rate is higher for other encapsulations, as high as 250 Mpps.

The loss of some TCP ACK packets are not the primary concern when such a burst occurs. When a burst occurs, any other packets, regardless of packet length are dropped once input buffers are exceeded. Buffers in front of the packet decision engine are often very small.

Internet service providers and content providers generally specify

full rate forwarding with 40 byte payload packets as a requirement. This requirement often can be waived if the provider can be convinced that when long sequence of short packets occur no packets will be dropped.

With adequate buffers before the packet decision engine, an LSR can absorb a long sequence of short packets. Even if the output is slowed to the point where light congestion occurs, the packets, having cleared the decision process, can make use of larger VOQ or output side buffers and be dealt with according to configured QoS treatment, rather than dropped completely at random.

Packet rate requirements apply regardless of which network tier equipment is deployed in. Whether deployed in the network core or near the network edges, packets must be processed at full line rate or with sufficient buffering prior to the packet decision engine.

2.3. MPLS Multipath Techniques

In any large provider, service providers and content providers, hash based multipath techniques are used in the core. In many of these providers hash based multipath is used in the edge as well and in some cases the metro.

The most common multipath techniques are ECMP applied at the IP forwarding level, Ethernet LAG with inspection of the IP payload, and multipath on links carrying both IP and MPLS, where the IP header is inspected below the MPLS label stack. In most core networks, the vast majority of traffic is MPLS encapsulated.

In order to support an adequately even load distribution across multiple links, IP addresses must be used. Common practice today is to reinspect the IP addresses at each LSR and use the label stack and IP addresses in a hash performed at each LSR.

The use of this technique is so ubiquitous in large core networks that lack of support for multipath makes any product unsuitable for use in large core networks. This will continue to be the case in the near future, even as deployment of MPLS Entropy Label begins to relax the core LSR multipath performance requirements given the existing deployed base of edge equipment without the ability to add an Entropy Label.

A generation of edge equipment supporting the ability to add an MPLS Entropy Label is needed before the performance requirements for core LSR can be relaxed. However, it is likely that two generations of deployment in the future will allow core LSR to support full packet rate only when a relatively small number of MPLS labels need to be

inspected before hashing. For now, don't count on it.

2.3.1. Pseudowire Control Word

Within the core of a network some form of multipath is almost certain to be used. Multipath techniques deployed today are likely to be looking beneath the label stack for an opportunity to hash on IP addresses.

A pseudowire encapsulated at a network edge must have a means to prevent reordering within the core if the pseudowire will be crossing a network core, or any part of a network topology where multipath is used.

Not supporting the ability to encapsulate a pseudowire with a control word may lock a product out from consideration. A pseudowire capability without control word support might be sufficient for applications which are strictly both intra-metro and low bandwidth. However a provider with other applications will very likely not tolerate having equipment which can only support a subset of their pseudowire needs.

2.3.2. Pseudowire Flow Label

Unlike a pseudowire control word, a pseudowire flow label [RFC6391], is required only for relatively large capacity pseudowires. There are many cases where a pseudowire flow label makes sense. Any service such as a VPN which carries IP traffic within a pseudowire can make use of a pseudowire flow label.

Any pseudowire which does not carry a flow label is in effect a single microflow (in [RFC2475] terms). Where multipath makes use of a simple hash (see Section 2.3) the presense of large microflows that consumes 10% of the capacity of a potentially congested link, can upset the traffic balance and in effect reduce the effective capacity of the entire microflow by far more than 10%. Therefore is a network where a significant number of parallel 10 Gb/s links exists, even a 1 Gb/s pseudowire should carry a flow label if possible.

2.3.3. MPLS Entropy Label

The MPLS Entropy Label simplifies flow group identification [I-D.ietf-mpls-entropy-label] in the network core. Prior to the MPLS Entropy Label core LSR needed to inspect the entire label stack and often the IP headers to provide an adequate distribution of traffic when using multipath techniques (see Section 2.3). With the use of MPLS Entropy Label, a hash can be performed closer to network edges, placed in the label stack, and used within the network core.

The MPLS Entropy Label avoid full label stack and payload inspection within the core where performance levels are most difficult to achieve (see Section 2.2). The label stack inspection can be terminated as soon as the first Entropy Label is encountered, which is generally after a small number of labels are inspected.

In order to provide these benefits in the core, LSR closer to the edge must be capable of adding an entropy label. This support may not be required in the access tier, the tier closest to the customer, but is likely to be required in the edge or the border to the network core. LSR peering with external networks will also need to be able to add an Entropy Label.

2.4. MPLS-TP and UHP

MPLS-TP introduces forwarding demands that will be extremely difficult to meet in a core network. Most troublesome is the requirement for Ultimate Hop Popping (UHP, the opposite of Penultimate Hop Popping or PHP). Using UHP opens the possibility of one or more MPLS POP operation plus an MPLS SWAP operation for each packet. The potential for multiple lookups and multiple counter instances per packet exists.

As networks grow and tunneling of LDP LSPs into RSVP-TE LSPs is used, and/or RSVP-TE hierarchy is used, the requirement to perform one or two or more MPLS POP operations plus a MPLS SWAP operation (and possibly a PUSH or two) increases. If MPLS-TP LM (link monitoring) OAM is enabled at each layer, then a packet and byte count must be maintained for each POP and SWAP operation.

3. Questions for Suppliers

The following questions should be asked of a supplier. These questions are grouped into broad categories.

Basic Compliance

- Q#1 Can the implementation forward packets with an arbitrarily large stack depth?

Basic Performance

- Q#2 Can very small packets be forwarded at full line rate on all interfaces indefinitely?

- Q#3 Customers must decide whether to relax the prior requirement and to what extent. If the answer to the prior question is "no", then:
- a. What is the smallest packet size where full line rate forwarding can be supported?
 - b. What is the longest burst of full rate small packets that can be supported?
- Q#4 How many POP operations can be supported along with a SWAP operation at full line rate while maintaining per LSP packet and byte counts for each POP and SWAP? This requirement is particularly relevant for MPLS-TP.
- Q#5 For a worst case where all packets arrive on one LSP, what is the counter overflow time? Are any means provided to avoid polling all counters at short intervals? This applies to both MPLS and MPLS-TP.

Multipath Capabilities and Performance

Multipath capabilities do not apply to MPLS-TP but apply to MPLS and apply if MPLS-TP is carried in MPLS.

- Q#6 How many MPLS labels can be included in a hash based on the MPLS label stack?
- Q#7 Is packet rate performance decreased beyond some number of labels?
- Q#8 Can the IP addresses below the MPLS stack be used in the hash?
- Q#9 At what maximum MPLS label stack depth can Bottom of Stack and an IP header appear without impacting packet rate performance?
- Q#10 Are reserved labels included in the label stack hash? They MUST NOT be included.

Pseudowire Capabilities and Performance

- Q#11 Is the pseudowire control word supported?

- Q#12 What is the maximum rate of pseudowire encapsulation and decapsulation? Apply the same questions as in Based Performance for any packet based pseudowire such as IP VPN or Ethernet.
- Q#13 Does inclusion of a pseudowire control word impact performance?
- Q#14 Are flow labels supported?
- Q#15 If so, what fields are hashed on for the flow label for different types of pseudowires?
- Q#16 Does inclusion of a flow label impact performance?

Entropy Label Support and Performance

- Q#17 Can an entropy label be added when acting as an ingress LER and can it be removed when acting as an egress LER?
- Q#18 If so, what fields are hashed on for the entropy label?
- Q#19 Does adding or removing an entropy label impact packet rate performance?
- Q#20 Can an entropy label be detected in the label stack, used in the hash, and properly terminate the search for further information to hash on?
- Q#21 Does using an entropy label have any negative impact on performance? It should have no impact or a positive impact.

4. Forwarding Compliance and Performance Testing

Packet rate performance of equipment supporting a large number of 10 Gb/s or 100 Gb/s links is not possible using desktop computers or workstations. The use of high end workstations as a source of test traffic was barely viable 20 years ago, but is no longer at all viable. Though custom microcode has been used on specialized router forwarding cards to serve the purpose of generating test traffic and measuring it, for the most part performance testing will require specialized test equipment. There are multiple sources of suitable equipment.

The set of tests listed here do not correspond one-to-one to the set of questions in Section 3. The same categorization is used and these

tests largely serve to validate answers provided the the prior questions, and can also provide answers where a supplier is unwilling to disclose compliance or performance.

Performance testing is the domain of the IETF Benchmark Methodology Working Group (BMWG). Below are brief descriptions of conformance and performance tests. Some very basic tests are specified in [RFC5695] which partially cover only the basic performance test T#2.

The following tests should be performed by the systems designer, or deployer, or performed by the supplier on their behalf if it is not practical for the potential customer to perform the tests directly. These tests are grouped into broad categories.

Basic Compliance

- T#1 Test forwarding at a high rate for packets with varying number of label entriess. While packets with more than a dozen label entriess are unlikely to be used in any practical scenario today, it is useful to know if limitations exists.

Basic Performance

- T#2 Test packet forwarding at full line rate with small packets. See [RFC5695]. The most likely case to fail is the smallest packet size.
- T#3 If the prior tests did not succeed for all packat sizes, then perform the following tests.
 - a. Increase the packet size by 4 bytes until a size is found that can be forwarded at full rate.
 - b. Inject bursts of consecutive small packets into a stream of larger packets. Allow some time for recovery between bursts. Increase the number of packets in the burst until packets are dropped. One way to accomplish this is to use a router with higher priority set on the interfaces on which small packets are sent to it. The router should buffer the lower priority large packets. It is best to inject the small packets to this router on a faster interface (if such a thing exists), or more than one interface.

- T#4 Send test traffic where a SWAP operation is required. Also set up multiple LSP carried over other LSP where the device under test (DUT) is the egress of these LSP. Create test packets such that the SWAP operation is performed after POP operations, increasing the number of POP operations until forwarding of small packets at full line rate can no longer be supported. Also check to see at what point the full set of counters can no longer be maintained. This requirement is particularly relevant for MPLS-TP.
- T#5 Send all traffic on one LSP and see if the counters become inaccurate. Often counters on silicon are much smaller than the 64 bit counters in IETF MIB. System developers should consider what counter polling rate is necessary to maintain accurate counters and whether those polling rates are practical. Relevant MIBs for MPLS are discussed in [RFC4221] and [RFC6639].

Multipath Capabilities and Performance

Multipath capabilities do not apply to MPLS-TP but apply to MPLS and apply if MPLS-TP is carried in MPLS.

- T#6 Send traffic at a rate well exceeding the capacity of a single multipath component link, and where entropy exists only below the top of stack. If only the top label is used this test will fail immediately.
- T#7 Move the labels with entropy down in the stack until either the full forwarding rate can no longer be supported or most or all packets try to use the same component link.
- T#8 Repeat the two tests above with the entropy contained in IP addresses below the label stack rather than in the label stack.
- T#9 Determine whether traffic that contains a pseudowire control word is interpreted as IP traffic. Information in the payload MUST NOT be used in the load balancing if the first nibble of the packet is not 4 or 6 (IPv4 or IPv6).
- T#10 Determine whether reserved labels are included in the label stack hash. They MUST NOT be included.

Pseudowire Capabilities and Performance

- T#11 Determine whether pseudowire can be set up with a pseudowire label and pseudowire control word added at ingress and the pseudowire label and pseudowire control word removed at egress.
- T#12 For pseudowire that contains variable length payload packets, repeat the packet size based performance tests for pseudowire ingress and egress functions.
- T#13 Repeat pseudowire performance tests with and without a pseudowire control word.
- T#14 Determine whether pseudowire can be set up with a pseudowire label, flow label, and pseudowire control word added at ingress and the pseudowire label, flow label, and pseudowire control word removed at egress.
- T#15 Determine which payload fields are used to create the flow label and whether the set of fields and algorithm provide sufficient entropy for load balancing.
- T#16 Repeat pseudowire performance tests with flow labels included.

Entropy Label Support and Performance

- T#17 Determine whether entropy labels are supported.
- T#18 Determine which fields are used to create an entropy label. Labels further down in the stack, including entropy labels further down and IP payload where applicable should be used. Determine whether the set of fields and algorithm provide sufficient entropy for load balancing.
- T#19 Repeat performance tests at LSP ingress and egress when entropy labels are added or removed.
- T#20 Determine whether an ELI is detected when acting as a midpoint LSR and whether the search for further information on which to base the load balancing is used. Information below the entropy label MUST NOT be used.
- T#21 Repeat performance tests for midpoint LSR with entropy labels found at various label stack depths.

5. IANA Considerations

This memo includes no request to IANA.

6. Security Considerations

This document reviews forwarding behaviour specified elsewhere and points out compliance and performance requirements. As such it introduces no new security requirements or concerns. Knowledge of potential performance shortcomings may serve to help avoid pitfalls, but in very unlikely circumstances such knowledge could in principle be the basis of denial of service. In practice such extreme data and packet rate would be needed to make this type of denial of service extremely unlikely and undetectable denial of service impossible.

7. References

7.1. Normative References

- [I-D.ietf-mpls-entropy-label] Kompella, K., Drake, J., Amante, S., Henderickx, W., and L. Yong, "The Use of Entropy Labels in MPLS Forwarding", draft-ietf-mpls-entropy-label-06 (work in progress), September 2012.
- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC3032] Rosen, E., Tappan, D., Fedorkow, G., Rekhter, Y., Farinacci, D., Li, T., and A. Conta, "MPLS Label Stack Encoding", RFC 3032, January 2001.
- [RFC3209] Awduche, D., Berger, L., Gan, D., Li, T., Srinivasan, V., and G. Swallow, "RSVP-TE: Extensions to RSVP for LSP Tunnels", RFC 3209, December 2001.
- [RFC3270] Le Faucheur, F., Wu, L., Davie, B., Davari, S., Vaananen, P., Krishnan, R., Cheval, P., and J. Heinanen, "Multi-Protocol Label Switching (MPLS) Support of Differentiated Services", RFC 3270, May 2002.
- [RFC3443] Agarwal, P. and B. Akyol, "Time To Live (TTL) Processing in Multi-Protocol Label Switching (MPLS) Networks", RFC 3443, January 2003.
- [RFC4090] Pan, P., Swallow, G., and A. Atlas, "Fast Reroute

Extensions to RSVP-TE for LSP Tunnels", RFC 4090, May 2005.

- [RFC4182] Rosen, E., "Removing a Restriction on the use of MPLS Explicit NULL", RFC 4182, September 2005.
- [RFC4201] Kompella, K., Rekhter, Y., and L. Berger, "Link Bundling in MPLS Traffic Engineering (TE)", RFC 4201, October 2005.
- [RFC5129] Davie, B., Briscoe, B., and J. Tay, "Explicit Congestion Marking in MPLS", RFC 5129, January 2008.
- [RFC5586] Bocci, M., Vigoureux, M., and S. Bryant, "MPLS Generic Associated Channel", RFC 5586, June 2009.
- [RFC6391] Bryant, S., Filsfils, C., Drafz, U., Kompella, V., Regan, J., and S. Amante, "Flow-Aware Transport of Pseudowires over an MPLS Packet Switched Network", RFC 6391, November 2011.

7.2. Informative References

- [ACK-compression] "Observations and Dynamics of a Congestion Control Algorithm: The Effects of Two-Way Traffic", Proc. ACM SIGCOMM, ACM Computer Communications Review (CCR) Vol 21, No 4, 1991, pp.133-147., 1991.
- [RFC2475] Blake, S., Black, D., Carlson, M., Davies, E., Wang, Z., and W. Weiss, "An Architecture for Differentiated Services", RFC 2475, December 1998.
- [RFC3031] Rosen, E., Viswanathan, A., and R. Callon, "Multiprotocol Label Switching Architecture", RFC 3031, January 2001.
- [RFC4124] Le Faucheur, F., "Protocol Extensions for Support of Diffserv-aware MPLS Traffic Engineering", RFC 4124, June 2005.
- [RFC4206] Kompella, K. and Y. Rekhter, "Label Switched Paths (LSP) Hierarchy with Generalized Multi-Protocol Label Switching (GMPLS) Traffic Engineering (TE)", RFC 4206, October 2005.
- [RFC4221] Nadeau, T., Srinivasan, C., and A. Farrel, "Multiprotocol Label Switching (MPLS) Management Overview", RFC 4221, November 2005.
- [RFC5462] Andersson, L. and R. Asati, "Multiprotocol Label Switching

(MPLS) Label Stack Entry: "EXP" Field Renamed to "Traffic Class" Field", RFC 5462, February 2009.

[RFC5695] Akhter, A., Asati, R., and C. Pignataro, "MPLS Forwarding Benchmarking Methodology for IP Flows", RFC 5695, November 2009.

[RFC6639] King, D. and M. Venkatesan, "Multiprotocol Label Switching Transport Profile (MPLS-TP) MIB-Based Management Overview", RFC 6639, June 2012.

Authors' Addresses

Curtis Villamizar (editor)
Outer Cape Cod Network Consulting

Email: curtis@occnc.com

Kireeti Kompella
Contrail Systems

Email: kireeti.kompella@gmail.com

MPLS
Internet-Draft
Intended status: Informational
Expires: April 5, 2013

C. Villamizar, Ed.
Outer Cape Cod Network
Consulting
October 2, 2012

Use of Multipath with MPLS-TP and MPLS
draft-villamizar-mpls-tp-multipath-03

Abstract

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

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

Status of this Memo

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

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

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

This Internet-Draft will expire on April 5, 2013.

Copyright Notice

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

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

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

Table of Contents

| | |
|---|---|
| 1. Introduction | 3 |
| 2. Definitions | 3 |
| 3. MPLS as a Server Layer for MPLS-TP | 5 |
| 4. MPLS-TP as a Server Layer for MPLS | 7 |
| 5. IANA Considerations | 8 |
| 6. Security Considerations | 8 |
| 7. References | 8 |
| 7.1. Normative References | 8 |
| 7.2. Informative References | 8 |
| Author's Address | 9 |

1. Introduction

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

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

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

2. Definitions

Multipath

The term multipath includes all techniques in which

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

- c. may be links or paths to an intermediate node used to reach a common destination.

Link Bundle

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

Link Aggregation

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

Link Aggregation Group (LAG)

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

Equal Cost Multipath (ECMP)

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

Loop Free Alternate Paths

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

Composite Link

The term Composite Link had been a registered trademark of Avici Systems, but was abandoned in 2007. The term composite link is now defined by the ITU in [ITU-T.G.800]. The ITU definition

includes multipath as defined here, plus inverse multiplexing which is explicitly excluded from the definition of multipath.

Inverse Multiplexing

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

Component Link

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

LAG Member

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

load split

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

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

3. MPLS as a Server Layer for MPLS-TP

MPLS LSP may be used as a server layer for MPLS-TP LSP as long as all MPLS-TP requirements are met, including the requirement that packets within an MPLS-TP LSP are not reordered, including both payload and OAM packets.

Supporting MPLS-TP LSP over a fully MPLS-TP conformant MPLS LSP server layer where the MPLS LSP are making use of multipath, requires special treatment of the MPLS-TP LSP such that those LSP only are not subject to the multipath load splitting. This implies the following brief set of requirements.

- MP#1 It MUST be possible to identify MPLS-TP LSP.
- MP#2 It MUST be possible to completely exclude MPLS-TP LSP from the multipath hash and load split.
- MP#3 It SHOULD be possible to insure that an MPLS-TP LSP will not be moved to another component link as a result of a composite link load rebalancing operation.
- MP#4 Where an RSVP-TE control plane is used, it MUST be possible for an ingress LSR which is setting up an MPLS-TP or MPLS LSP to determine at CSPF time whether a link or MPLS PSC LSP within the topology can support the MPLS-TP requirements of the LSP.

There is currently no signaling mechanism defined to support requirement MP#1. In the absence of a signaling extension, MPLS-TP can be identified through some form of configuration, such as configuration which provides an MPLS-TP compatible server layer to all LSP arriving on a specific interface or originating from a specific set of ingress LSR. Alternately an MPLS-TP LSP can be created with an Entropy Label Indicator (ELI) and entropy label (EL) below the MPLS-TP label [I-D.ietf-mpls-entropy-label].

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

MPLS-TP LSP can be carried as client LSP within an MPLS server LSP if an Entropy Label Indicator (ELI) and entropy label (EL) is added after the server layer LSP label(s) in the label stack, just above the MPLS-TP LSP label entry [I-D.ietf-mpls-entropy-label]. This allows MPLS-TP LSP to be carried as client LSP within MPLS LSP and satisfies requirement MP#2 but requires that MPLS LSR be able to identify MPLS-TP LSP (requirement MP#1).

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

If MPLS-TP LSP are carried within MPLS LSP and ELI and EL are used,

requirement MP#2 is satisfied, but without a signaling extension, requirement MP#3 is not satisfied if there is a need to rebalance the load on any composite link carrying the MPLS server LSP. Load rebalance is generally needed only when congestion occurs, therefore restricting MPLS-TP to be carried only over MPLS LSP that are known to traverse only links which are expected to be uncongested can satisfy requirement MP#3.

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

4. MPLS-TP as a Server Layer for MPLS

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

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

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

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

discussed above.

5. IANA Considerations

This memo includes no request to IANA.

6. Security Considerations

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

7. References

7.1. Normative References

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

7.2. Informative References

[I-D.ietf-mpls-entropy-label]
Kompella, K., Drake, J., Amante, S., Henderickx, W., and L. Yong, "The Use of Entropy Labels in MPLS Forwarding", draft-ietf-mpls-entropy-label-06 (work in progress), September 2012.

[I-D.ietf-mpls-tp-security-framework]
Fang, L., Niven-Jenkins, B., Mansfield, S., and R. Graveman, "MPLS-TP Security Framework", draft-ietf-mpls-tp-security-framework-04 (work in progress), July 2012.

[IEEE-802.1AX]
IEEE Standards Association, "IEEE Std 802.1AX-2008 IEEE Standard for Local and Metropolitan Area Networks - Link Aggregation", 2006, <<http://standards.ieee.org/getieee802/download/802.1AX-2008.pdf>>.

[ITU-T.G.800]

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

- [RFC2475] Blake, S., Black, D., Carlson, M., Davies, E., Wang, Z., and W. Weiss, "An Architecture for Differentiated Services", RFC 2475, December 1998.
- [RFC3209] Awduche, D., Berger, L., Gan, D., Li, T., Srinivasan, V., and G. Swallow, "RSVP-TE: Extensions to RSVP for LSP Tunnels", RFC 3209, December 2001.
- [RFC4201] Kompella, K., Rekhter, Y., and L. Berger, "Link Bundling in MPLS Traffic Engineering (TE)", RFC 4201, October 2005.
- [RFC5286] Atlas, A. and A. Zinin, "Basic Specification for IP Fast Reroute: Loop-Free Alternates", RFC 5286, September 2008.
- [RFC5462] Andersson, L. and R. Asati, "Multiprotocol Label Switching (MPLS) Label Stack Entry: "EXP" Field Renamed to "Traffic Class" Field", RFC 5462, February 2009.
- [RFC5654] Niven-Jenkins, B., Brungard, D., Betts, M., Sprecher, N., and S. Ueno, "Requirements of an MPLS Transport Profile", RFC 5654, September 2009.
- [RFC5714] Shand, M. and S. Bryant, "IP Fast Reroute Framework", RFC 5714, January 2010.
- [RFC5920] Fang, L., "Security Framework for MPLS and GMPLS Networks", RFC 5920, July 2010.

Author's Address

Curtis Villamizar (editor)
Outer Cape Cod Network Consulting
Email: curtis@ocnc.com

MPLS
Internet-Draft
Intended status: Standards Track
Expires: April 10, 2013

C. Villamizar, Ed.
Outer Cape Cod Network
Consulting
October 7, 2012

Multipath Extensions for MPLS Traffic Engineering
draft-villamizar-mpls-tp-multipath-te-extn-02

Abstract

Extensions to OSPF-TE, ISIS-TE, and RSVP-TE are defined in support of carrying LSP with strict packet ordering requirements over multipath and carrying LSP with strict packet ordering requirements within LSP without violating requirements to maintain packet ordering. LSP with strict packet ordering requirements include MPLS-TP LSP.

OSPF-TE and ISIS-TE extensions defined here indicate node and link capability regarding support for ordered aggregates of traffic, multipath traffic distribution, and abilities to support multipath load distribution differently per LSP.

RSVP-TE extensions either identifies an LSP as requiring strict packet order, or identifies an LSP as carrying one or more LSP that requires strict packet order further down in the label stack, or identifies an LSP as having no restrictions on packet ordering except the restriction to avoid reordering microflows. In addition an extension indicates whether the first nibble of payload will reliably indicate whether payload is IPv4, IPv6, or other type of payload, most notably pseudowire using a pseudowire control word.

Status of this Memo

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

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

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

This Internet-Draft will expire on April 10, 2013.

Copyright Notice

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

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

Table of Contents

| | |
|---|----|
| 1. Introduction | 4 |
| 1.1. Architecture Summary | 4 |
| 1.2. Requirements Language | 5 |
| 1.3. Definitions | 5 |
| 2. Protocol Extensions | 6 |
| 2.1. Multipath Node Capability sub-TLV | 6 |
| 2.2. Multipath Link Capability TLV | 9 |
| 2.3. LSP Multipath Attributes TLV | 9 |
| 3. Protocol Mechanisms | 12 |
| 3.1. OSPF-TE and ISIS-TE Advertisement | 12 |
| 3.1.1. Node Capability Advertisement | 12 |
| 3.1.2. Link Capability Advertisement | 12 |
| 3.1.3. Setting Max Depth and IP Depth | 12 |
| 3.1.4. Advertising Multipath as Link Bundling | 13 |
| 3.1.5. Hierarchical LSP Link Advertisement | 13 |
| 3.1.6. Advertisement of Legacy Multipath | 14 |
| 3.2. RSVP-TE LSP Attributes | 15 |
| 3.2.1. LSP Contained Ordered Aggregates Flags | 15 |
| 3.2.2. LSP Attributes for Ordered Aggregates | 17 |
| 3.2.3. Attributes for LSP without Packet Ordering | 17 |
| 3.3. Path Computation Constraints | 20 |
| 3.3.1. Link Multipath Capabilities and Path Computation | 20 |
| 3.3.1.1. Path Computation with Ordering Constraints | 20 |
| 3.3.1.2. Path Computation with No Ordering Constraint | 21 |
| 3.3.1.3. Path Computation for MPLS containing MPLS-TP | 21 |
| 3.3.2. Link IP Capabilities and Path Computation | 21 |
| 3.3.2.1. LSP without Packet Ordering Requirements | 22 |
| 3.3.2.2. LSP with Ordering Requirements | 22 |
| 3.3.3. Link Depth Limitations and Path Computation | 23 |
| 4. Backwards Compatibility | 24 |
| 4.1. Legacy Multipath Behavior | 24 |
| 4.2. Networks without Multipath Extensions | 24 |
| 4.2.1. Networks with MP Capability on all Multipath | 24 |
| 4.2.2. Networks with OA Capability on all Multipath | 26 |
| 4.2.3. Legacy Networks with Mixed MP and OA Links | 26 |
| 4.3. Transition to Multipath Extension Support | 27 |
| 4.3.1. Simple Transitions | 27 |
| 4.3.2. More Challenging Transitions | 27 |
| 5. IANA Considerations | 28 |
| 6. Security Considerations | 28 |
| 7. References | 28 |
| 7.1. Normative References | 28 |
| 7.2. Informative References | 29 |
| Author's Address | 30 |

1. Introduction

Today the requirement to handle large aggregations of traffic, can be handled by a number of techniques which we will collectively call multipath. Multipath is very similar to composite link as defined in [ITU-T.G.800], except multipath specifically excludes inverse multiplexing. Some types of LSP, including but potentially not limited to MPLS-TP LSP, require strict packet ordering.

A means to support a MPLS-TP client layer over a MPLS server layer using MPLS Entropy Label is defined in [I-D.villamizar-mpls-tp-multipath]. It is noted in [I-D.villamizar-mpls-tp-multipath] that absent some protocol extensions, some limitations must be accepted.

This document defines protocol extensions which better supports using MPLS with multipath as a server layer for MPLS-TP, or to carry MPLS-TP directly over a network which makes use of multipath. Extensions are also applicable to MPLS when used in the presense of very large microflows or very large LSP which cannot be load split as a result of using MPLS Entropy Label [I-D.ietf-mpls-entropy-label].

1.1. Architecture Summary

Advertisements in a link state routing protocol, such as OSPF or ISIS, support a topology map known as a link state database (LSDB). When traffic engineering information is included in the LSDB the topology map is known as a TE-LSDB or traffic engineering database (TED).

A common MPLS LSP path computation is known as a constrained shortest path first computation (CSPF) (see [RFC3945]). Other algorithms may be used for path computation. Constraint-based routing was first introduced in [RFC2702]).

OSPF-TE or ISIS-TE extensions are defined in Section 2.1 and Section 2.2. OSPF-TE or ISIS-TE advertisements serve to populate the TE-LSDB and provide the basis for constraint-based routing path computation. Section 3.1 describes the use of OSPF-TE or ISIS-TE multipath extensions in routing advertisements.

RSVP-TE extensions are defined in Section 2.3. Section 3.2 describes the use of RSVP-TE extensions in setting up LSP including signaling constraints on LSP which contain other LSP which specify RSVP-TE extensions.

Section 3.3 describes the constraints on LSP path computation imposed by the advertised ordered aggregate and multipath capabilities of

links. Section 3.3.2 describes the constraints on LSP path computation imposed by link advertisements regarding use of IP headers in multipath traffic distribution. Section 3.3.3 describes the impact of label stack depth limitations.

1.2. Requirements Language

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

1.3. Definitions

Please refer to [I-D.villamizar-mpls-tp-multipath].

Ordered Aggregate (OA)

An ordered aggregate (OA) requires that packets be delivered in the order in which they were received. Please refer to [RFC3260].

Microflow

A microflow is a single instance of an application-to-application flow. Please refer to [RFC2475]. Reordering packets within a microflow can cause service disruption. Please refer to [RFC2991].

Multipath Traffic Distribution

Multipath traffic distribution refers to the mechanism which distributes traffic among a set of component links or component lower layer paths which together comprise a multipath. No assumptions are made about the algorithms used in multipath traffic distribution. This document only discusses constraints of the type of information which can be used as the basis for multipath traffic distribution in specific circumstances.

The phrase "strict packet ordering requirements" refers to the requirement to deliver all packet in the order that they were received. The absence of strict packet ordering requirements does not imply total absence of packet ordering requirements. The requirement to avoid reordering traffic within any given microflow, as described in [RFC2991] applies to all traffic aggregates including all MPLS LSP.

The abbreviations ELI and EL are Entropy Label Indicator and Entropy Label, as defined in [I-D.ietf-mpls-entropy-label].

2. Protocol Extensions

This section defined protocol extensions to OSPF-TE, ISIS-TE, and RSVP-TE. Use of these extensions is described in Section 3 and Section 4.

Two capability sub-TLV are added to two TLV that are used in both OSPF-TE and ISIS-TE. The Multipath Node Capability sub-TLV is added to the Node Attribute TLV (see Section 2.1). The Multipath Link Capability TLV is added to the Interface_ID (see Section 2.2).

One TLV is added to the LSP_ATTRIBUTES object defined in [RFC5420].

2.1. Multipath Node Capability sub-TLV

The Node Attribute TLV is defined in [RFC5786]. A new sub-TLV, the Multipath Node Capability sub-TLV, is defined for inclusion in the Node Attribute TLV.

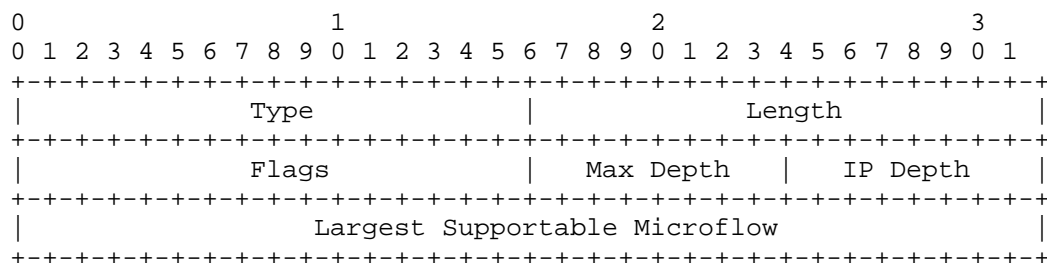


Figure 1: Multipath Capability Sub-TLV

The fields in the Multipath Capability sub-TLV are defined as follows.

Type

The Type field is assigned a value of IANA-TBD-1. The Type field is a two octet value.

Length

The Length field indicates the length of the sub-TLV in octets, excluding the Type and Length fields. The Length field is a two octet value.

Flags

The Flags field is a two octet (16 bit) value. The following single bit fields are assigned within this value, starting at the most significant bit, which is the bit transmitted first.

0x8000 Ordered Aggregate Enabled

Setting the Ordered Aggregate Enabled bit indicates that an LSP can be carried as an Ordered Aggregate Enabled on one or more links.

0x4000 Multipath Enabled

Setting the Multipath Enabled bit indicates that an LSP can be spread across component links at one or more multipath links.

0x2000 IPv4 Enabled Multipath

Setting the IPv4 Enabled Multipath bit indicates that the IPv4 header information can be used in multipath load balance. The Multipath Enabled bit must be set if the IPv4 Enabled Multipath bit is set.

0x1000 IPv6 Enabled Multipath

Setting the IP bit indicates that the IPv6 header information can be used in multipath load balance. The Multipath Enabled bit must be set if the IPv6 Enabled Multipath bit is set.

0x0800 UDP/IPv4 Multipath

Setting the UDP/IPv4 Multipath bit indicates that the UDP port numbers carried in UDP over IPv4 can be used in multipath load balance. The IPv4 Enabled Multipath bit must be set if UDP/IPv4 Multipath is set. If the IPv4 Enabled Multipath bit is set and the UDP/IPv4 Multipath bit is clear, then only source and destination IP addresses are used.

0x0400 UDP/IPv6 Multipath

Setting the UDP/IPv6 Multipath bit indicates that the UDP port numbers carried in UDP over IPv6 can be used in multipath load balance. The IPv6 Enabled Multipath bit must be set if UDP/IPv6 Multipath is set. The IPv6 Enabled Multipath bit must be set if UDP/IPv6 Multipath is set. If the IPv6 Enabled Multipath bit is set and the UDP/IPv6 Multipath bit is clear, then only source and destination IP addresses are used.

0x0200 TCP/IPv4 Multipath

Setting the TCP/IPv4 Multipath bit indicates that the TCP port numbers carried in TCP over IPv4 can be used in multipath load balance. The IPv4 Enabled Multipath bit must be set if TCP/IPv4 Multipath is set. If the IPv4 Enabled Multipath bit is set and the TCP/IPv4 Multipath bit is clear, then only source and destination IP addresses are used.

0x0100 TCP/IPv6 Multipath

Setting the TCP/IPv6 Multipath bit indicates that the TCP port numbers carried in TCP over IPv6 can be used in multipath load balance. The IPv6 Enabled Multipath bit must be set if TCP/IPv6 Multipath is set. The IPv6 Enabled Multipath bit must be set if TCP/IPv6 Multipath is set. If the IPv6 Enabled Multipath bit is set and the TCP/IPv6 Multipath bit is clear, then only source and destination IP addresses are used.

0x0080 Default to Multipath

Setting the Default to Multipath bit indicates that for an LSP which does not signal a desired behavior the traffic for that LSP will be spread across component links at one or more multipath links. If the Default to Multipath bit is not set, then an LSP which does not signal otherwise will be treated as an ordered aggregate.

0x0040 Default to IP/MPLS Multipath

Setting the Default to IP/MPLS Multipath indicates that for an LSP which does not signal a desired behavior, the IP header information will be used in the multipath load distribution. If the Default to IP/MPLS Multipath is clear it indicates that the the IP header information will not be used by default.

0x0020 Entropy Label Multipath

Setting the Entropy Label Multipath bit indicates that when multipath is enabled for a given LSP, if an Entropy Label Indicator (ELI) is found in the label stack, information below the Entropy Label (EL) will not be used in multipath load distribution.

0x0010 IP Optional Multipath

Setting the IP Optional Multipath bit indicates that when multipath is enabled for a given LSP, whether the IP header information is used in the multipath load distribution can be set on a per LSP basis.

The remaining bits in the Flags field are reserved.

Max Depth

The Max Depth field is a one octet field indicating the maximum label stack depth beyond which the multipath load distribution cannot make use of further label stack entries.

IP Depth

The IP Depth field is a one octet field indicating the maximum label stack depth beyond which the multipath load distribution cannot make use of IP information.

Largest Supportable Microflow

The Largest Supportable Microflow field is a IEEE 32 bit floating point value expressing in bytes/second. Any microflow which exceeds this capacity may experience either packet loss, or out-of-order delivery, or both.

The reserved bits in the Flags field MUST be set to zero and MUST be ignored unless implementing an extension which redefines one or more of the reserved bits. Any further extension which redefines one or more reserved Flags bit should maintain backwards compatibility with prior implementations.

2.2. Multipath Link Capability TLV

The Interface_ID is defined in [RFC3471]. The Interface_ID is updated in [RFC4201] to support a form of multipath known as Link Bundling.

A new TLV, the Multipath Link Capability TLV, is defined here. The Multipath Link Capability TLV is optionally included in the Interface_ID. The format of the Multipath Link Capability TLV is identical to the Multipath Node Capability sub-TLV defined in Section 2.1, with one exception. In the Multipath Link Capability TLV the Type field value is IANA-TBD-2.

If a Multipath Link Capability TLV is advertised for any link, then a Multipath Node Capability sub-TLV MUST be advertised for the node.

2.3. LSP Multipath Attributes TLV

The LSP_ATTRIBUTES object is defined in [RFC5420]. A new LSP Multipath Attributes TLV is defined for the LSP_ATTRIBUTES object. The TLV Type is IANA_TBD_3. The format is described below.

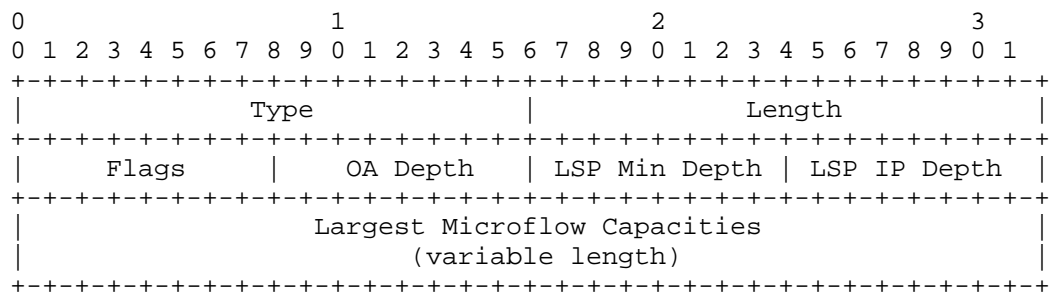


Figure 2: LSP Multipath Attributes TLV

The fields in the LSP Multipath Attributes TLV are defined as

follows.

Type

The Type field is assigned a value of IANA-TBD-3. The Type field is a two octet value.

Length

The Length field indicates the length of the sub-TLV in octets, excluding the Type and Length fields. The Length field is a two octet value.

Flags

The Flags field is a one octet (8 bit) value. The following single bit fields are assigned within this value, starting at the most significant bit, which is the bit transmitted first.

0x80 IP Multipath Allowed

Setting the IP Multipath Allowed bit indicates that it is safe to enable the use of a potential IP payload in the multipath traffic distribution.

0x40 May Contain IPv4

Setting the May Contain IPv4 bit indicates that IPv4 traffic may be contained within this LSP.

0x20 May Contain IPv6

Setting the May Contain IPv6 bit indicates that IPv6 traffic may be contained within this LSP.

0x02 Entropy Label Required

Setting the Entropy Label Used bit indicates that midpoint LSR MUST support ELI and EL in order to not violate packet ordering constraints of the LSP or of contained LSP.

0x01 Entropy Label Used

Setting the Entropy Label Used bit indicates that an ELI and EL is present in some or all label stack entries within this LSP.

The remaining bits in the Flags field are reserved.

OA Depth

The OA Depth field is set as follows

- 0 An OA Depth value of zero indicates that no ordered aggregates are carried within the LSP, except those which are protected from out of order delivery using Entropy Label.

- 1 An OA Depth value of one indicates that the LSP is an ordered aggregate of traffic (the LSP requires strict ordering of packets) and has protected packet ordering using ELI and EL.
- >1 An OA Depth value greater than one indicates that the LSP does not have strict packet ordering requirements but contains ordered aggregates at the label stack depth indicated or deeper and that the ordered aggregates are not protected using ELI and EL.

LSP Min Depth

The LSP Min Depth field indicates a minimal acceptable number of label used in multipath traffic distribution for the stated Largest Microflow Capacities field to be valid. If the LSP Min Depth field is set to zero this value is unknown. See Section 3.3.3.

LSP IP Depth

The LSP IP Depth field indicates a minimal label stack depth where using an IP header is necessary for the stated Largest Microflow Capacities field to be valid. If the LSP IP Depth field is set to zero this value is unknown. See Section 3.3.3.

Largest Microflow Capacities

The Largest Microflow Capacities field contains zero, one, two, or three IEEE 32 bit floating point values. Each value is a capacity expressed in bytes per second.

Largest LSE Microflow

The first value, the Largest LSE Microflow, is the capacity of the largest microflow if only the label stack entries are used in multipath traffic distribution. If a Largest LSE Microflow is not included, the LSP bandwidth request MUST be used.

Largest IP Microflow

The second value, the Largest IP Microflow, if present, is the capacity of the largest microflow if the label stack entries and any potential IP source and destination address are used in multipath traffic distribution. If the Largest IP Microflow is not included, the value of the Largest LSE Microflow MUST be used.

Largest L4 Microflow

The third, the Largest L4 Microflow, if present, is the capacity of the largest microflow if the label stack entries and any potential IP addresses and TCP or UDP port numbers are used in multipath traffic distribution. If a Largest L4 Microflow is not included, the value of the Largest IP

Microflow MUST be used.

3. Protocol Mechanisms

3.1. OSPF-TE and ISIS-TE Advertisement

Every compliant node MUST advertise exactly one Multipath Node Capability sub-TLV and MAY advertise zero or more Multipath Link Capability sub-TLV as needed.

Procedures for accommodating legacy forwarding capabilities and non-compliant nodes are discussed in Section 4.

3.1.1. Node Capability Advertisement

Every LSR which is adjacent to one or more multipath link MUST advertise a Multipath Node Capability sub-TLV (see Section 2.1). The capabilities advertised for the node SHOULD reflect the capabilities of the majority of multipath links adjacent to the node.

Every LSR which is not adjacent to any multipath links MUST advertise a Multipath Node Capability sub-TLV with both the Ordered Aggregate Enabled bit in Flags set and all other Flags bits clear. Both Max Depth and IP Depth can be set to zero. This advertisement identifies the LSR as one which is conformant but has no multipath links, allowing it to be distinguished from a non-conformant LSR with LAG or other multipath which may have to be avoided when determining a path for some LSP.

3.1.2. Link Capability Advertisement

For all of the links whose capability does not exactly match the Multipath Node Capability sub-TLV advertised by that same LSR, the LSR MUST advertise a Multipath Link Capability sub-TLV (see Section 2.2).

For all of the links whose capability does exactly match the Multipath Node Capability sub-TLV advertised by that same LSR, the LSR SHOULD NOT advertise a Multipath Link Capability sub-TLV (see Section 2.2). In this case the Multipath Link Capability TLV is redundant, but harmless.

3.1.3. Setting Max Depth and IP Depth

The Max Depth and IP Depth field are intended to capture architectural limits. Most forwarding hardware will only use a limited number of label entries in the multipath traffic

distribution. This limit is reflected in the Max Depth field. Most forwarding hardware will limit the number of label entries that it will look past before looking for an IP header to be used in the multipath traffic distribution. This limit is reflected in the IP Depth field.

3.1.4. Advertising Multipath as Link Bundling

All multipath links and FA for PSC LSP which traverse multipath links MAY be advertised as Link Bundles as defined in [RFC4201]. The settings of the Ordered Aggregate Enabled and Multipath Enabled fields indicate key capabilities of the multipath. Advertising the multipath as a link bundle can provide additional information, such as the capability to place LSP on individual components.

If the Multipath Enabled bit is set in the Multipath Link Capability TLV Flags, then the Maximum LSP Bandwidth in the Interface Identification TLV can be set in one of two ways.

1. If the desired behavior for legacy LSR acting as ingress is to limit LSP to the capacity of a single component link, then Maximum LSP Bandwidth SHOULD be set as specified in [RFC4201].
2. If the desired behavior for legacy LSR acting as ingress is to allow LSP to exceed the capacity of a single component link, then Maximum LSP Bandwidth MAY be set to a higher value, up to the value of Maximum Reservable Bandwidth. This would normally be done if the legacy LSR were known to be carrying traffic which is easily load split, such as typical Internet traffic.

LSR acting as ingress SHOULD ignore the Maximum LSP Bandwidth and MAY set up LSP with capacity up to the Maximum Reservable Bandwidth as long as microflows within the LSP will not exceed the Largest Supportable Microflow capacity.

3.1.5. Hierarchical LSP Link Advertisement

A PSC LSP, as defined in [RFC4206] and updated in [RFC6107], may carry other LSP. When signaling a PSC LSP expected characteristics of the contained traffic must be estimated. The FA advertised for the PSC LSP MUST reflect the broadest set of requirements the PSC LSP can carry. If the setup of an additional reservation would exceeded current capacity, a PSC LSP may be resigaled using make-before-break semantics ([RFC3209]).

For example, if it is expected that a PSC LSP will carry MPLS-TP LSP or other LSP with strict packet reordering requirements at some label depth, the minimum label stack depth at which an LSP with strict

packet reordering requirements can be carried must be signaled in the OA Depth field of the LSP Multipath Attributes TLV (see Section 2.3).

When the Forwarding Adjacency (FA) is advertised, the advertised Max Depth and IP Depth MUST be one less than the minimum of the Max Depth and IP Depth of any link that the PSC LSP traverses. The Max Depth and IP Depth are considered independently of each other. The reduction by one takes into account the PSC label. The Flags advertised for the FA MUST reflect the capabilities of the links along the path.

3.1.6. Advertisement of Legacy Multipath

An Ethernet LAG with no support for Entropy Label MUST have the Ordered Aggregate Enabled bit cleared and the Multipath Enabled bit set in the Multipath Link Capability TLV Flags. This indicates that a MPLS-TP compliant server layer cannot be supported and that LSP larger than the component links (LAG members) can be supported.

A Link Bundle that does not support the all-ones component defined in [RFC4201] MUST have the Ordered Aggregate Enabled bit set and the Multipath Enabled bit cleared in the Multipath Link Capability TLV Flags. This indicates that a MPLS-TP compliant server layer can be supported and that LSP larger than the component links cannot be supported.

A link bundle that can support either the pinning of a LSP to a single component link or the spreading of traffic across multiple component links MUST have the Ordered Aggregate Enabled bit set and the Multipath Enabled bit set in the Multipath Link Capability TLV Flags. This indicates that a MPLS-TP compliant server layer can be supported and that LSP larger than the component links can also be supported.

Any form of multipath that supports Entropy Label MUST have the Ordered Aggregate Enabled bit set and the Multipath Enabled bit set and the Entropy Label Multipath bit set in the Multipath Link Capability TLV Flags. Any form of multipath that does not support Entropy Label MUST have the Entropy Label Multipath bit cleared in the Multipath Link Capability TLV Flags.

The remaining bits in the Multipath Link Capability TLV Flags MUST be set according to the capability and configuration of the multipath or LSP.

3.2. RSVP-TE LSP Attributes

All LSR SHOULD advertise a LSP Multipath Attributes TLV with the RSVP-TE signaling for each LSP for which it is serving as ingress. If any legacy LSR remain in the network, it is easier to assign an acceptable default treatment for LSP signaled by those legacy LSR if the conforming LSR always send a LSP Multipath Attributes TLV.

There are two general cases, an LSP requires strict ordering of packets, or it doesn't. In the latter case the LSP may contain other LSP that require strict ordering and those must be protected from reordering using an Entropy Label as described in [I-D.villamizar-mpls-tp-multipath]. These two cases are briefly described below.

Ordered Aggregates

LSP with strict packet order requirements MUST set the OA Depth field to one to indicate that the LSP MUST be treated as ordered aggregate. See Section 3.2.2.

LSP without Packet Ordering

LSP which do not have strict packet order requirements MUST only carry LSP whose requirements are reflected in the containing LSP Multipath Attributes TLV. See Section 3.2.3.

If an attempt is made to signal a contained LSP whose Ordered Aggregate Attributes TLV and LSP Multipath Attributes TLV cannot be supported by the containing (PSC) LSP, one of the two following actions MUST be taken.

1. The containing (PSC) LSP MAY be resigaled with appropriate TLVs to carry the new contained LSP using make-before-break semantics, then continue to forward the contained LSP PATH message if the containing (PSC) LSP is successfully updated.
2. The LSR MAY reject the contained LSP signaling by sending a PathErr message.

3.2.1. LSP Contained Ordered Aggregates Flags

The Flags field in the LSP Multipath Attributes TLV MUST be set as follows.

1. If the LSP may directly contain IPv4 traffic, then the May Contain IPv4 bit in the Flags field MUST be set.
2. If the LSP may directly contain IPv6 traffic, then the May Contain IPv6 bit in the Flags field MUST be set.

3. If the LSP contains an LSP which has the May Contain IPv4 bit in the Flags field then the May Contain IPv4 bit in the Flags field MUST be set in the containing LSP.
4. If the LSP contains an LSP which has the May Contain IPv6 bit in the Flags field then the May Contain IPv6 bit in the Flags field MUST be set in the containing LSP.
5. If the LSP may contain pseudowires that do not use a pseudowire control word [RFC4385], and may contain IPv4 or IPv6 traffic, then the IP Multipath Allowed bit in the Flags field MUST be cleared.
6. If the LSP is known to contain no pseudowires that do not use a pseudowire control word, then the IP Multipath Allowed bit in the Flags field SHOULD be set unless disallowed due to a contained LSP.
7. If an Entropy Label is added below the LSP label, then the Entropy Label Used bit MUST be set.
8. If the LSP contains any LSP with the IP Multipath Allowed bit in the Flags field clear, then the IP Multipath Allowed bit in the Flags field MUST be clear.

If the LSP does not contain other LSP, it may contain IP traffic and/or pseudowire that terminate on that LSR. If the LSP does not contain other LSP. The LER will know whether the LSP is used in an IP LER capacity. The LER will also know whether it terminates any pseudowire for a given LSP. The LER will also know if it is using Entropy Label for a given LSP and if it requires strict packet ordering for a given LSP. Therefore, when a LSP does not contain other LSP, then it is possible to accurately set the Flags field in the LSP Multipath Attributes TLV, as well the OA Depth, and LSP IP Depth fields.

If an LSP contains other LSP, and all of the contained include a LSP Multipath Attributes TLV, then it is still possible to accurately set the Flags field in the LSP Multipath Attributes TLV, as well the OA Depth, and LSP IP Depth fields. It is only when LSP contains other LSP that do not have a LSP Multipath Attributes TLV where default behavior has to be configured based on assumptions about LSP originated by the legacy LSR where there is a potential for those configured assumptions to be inaccurate.

See Section 4 for guidelines for handling LSP which contain LSP that do not have a LSP Multipath Attributes TLV. The most conservative approach in this case is to clear the IP Multipath Allowed bit and

set the May Contain IPv4 bit and the May Contain IPv6 bit, however this may not always be necessary.

3.2.2. LSP Attributes for Ordered Aggregates

An LSP with strict packet order requirements MUST always include a LSP Multipath Attributes TLV.

Most of the Flags in the LSP Multipath Attributes TLV can be set as described in Section 3.2.1. There are two cases which affect the setting of the remaining Flags bits.

Entropy Label not used

If an Entropy Label is not used, then the Entropy Label Used bit, the Entropy Label Required bit, and the IP Multipath Allowed bit MUST be cleared.

Entropy Label is used If an Entropy Label is used, then the Entropy Label Used bit, and the Entropy Label Required bit, and the IP Multipath Allowed bit MUST be set.

The OA Depth field MUST be set to one. The Min Depth MUST be set to one. The LSP IP Depth SHOULD be set to zero. The Largest Microflow Capacities field SHOULD be empty. The entire LSP is one microflow. The Largest Microflow Capacities field MAY contain one entry if there is some reason to do so, such as an LSP which may peak at capacity higher than the bandwidth reserved for the LSP. The Largest Microflow Capacities for an LSP SHOULD be configurable independently of the LSP bandwidth reservation.

3.2.3. Attributes for LSP without Packet Ordering

If an LSP does not have strict packet order constraints, then the LSR_ATTRIBUTE object SHOULD always include a LSP Multipath Attributes TLV.

Most of the Flags in the LSP Multipath Attributes TLV can be set as described in Section 3.2.1. There are two cases which affect the setting of the remaining Flags bits, the OA Depth field, the LSP Min Depth, and the LSP IP Depth field.

Entropy Label not used

- * The OA Depth MUST be either set to zero or set to a configured value that is greater than one, or set based on the contained LSP.

- * If the OA Depth is set to a configured value, then any setup attempt for a contained LSP with a depth greater than or equal to that value SHOULD be rejected and a PathErr message sent. Otherwise, if a setup attempt for a contained LSP with a depth greater than the current value included in the containing LSP OA Depth field, then the containing LSP MUST be rerouted with a OA Depth field value greater than any of the contained OA Depth field values.
- * The Entropy Label Used bit MUST be set if any contained LSP has the Entropy Label Used bit set.
- * The Entropy Label Required bit MUST be set if any contained LSP has the Entropy Label Required bit set.

Entropy Label is used

- * The OA Depth MUST be set to zero.
- * The Entropy Label Used bit MUST be set.
- * The Entropy Label Required bit MUST be set if any contained LSP has the Entropy Label Required bit set.
- * The Entropy Label Required bit MUST be set if any contained LSP has the OA Depth field set to a non-zero value.
- * The Entropy Label Required bit SHOULD be clear if there are no contained LSP has the OA Depth field set to a non-zero value and no contained LSP with the Entropy Label Required bit set. In this case the Entropy Label Required bit MAY be set by configuration, generally in anticipation of needing it in the future to carry other LSP.
- * LSP Min Depth field MUST be set to three if the Entropy Label Required bit is set.
- * If the Entropy Label Required bit is not set, then the LSP Min Depth field and LSP IP Depth field SHOULD be set to three if there are no contained LSP. The LSP Min Depth field and LSP IP Depth MAY be configured to a minimum value greater than three, generally in anticipation of needing it in the future to carry other LSP.
- * If the Entropy Label Required bit is not set, and there are contained LSP, then the LSP Min Depth field MUST be set to a value greater than three.

- * If the Entropy Label Required bit is not set, the LSP Min Depth field MUST be set to a value that is at least the sum of three plus the LSP Min Depth field in any contained LSP.
- * If the Entropy Label Required bit is not set, and either the May Contain IPv4 bit or the May Contain IPv6 bit is set, then the LSP IP Depth field MUST be set to a value of one or greater.
- * If the Entropy Label Required bit is not set, and any contained LSP has the May Contain IPv4 bit or the May Contain IPv6 bit is set, then the LSP IP Depth field MUST be set to a value of two or greater.
- * If the Entropy Label Required bit is not set, and any contained LSP has the LSP IP Depth field set to a value greater than one, then the LSP IP Depth field MUST be set to a value greater than the highest LSP IP Depth value set in any contained LSP.

The values of the LSP Min Depth field and the LSP IP Depth field MAY be constrained to upper limits by configuration. If an attempt to setup a contained LSP would result in exceeding one of these limits, then the LSR SHOULD reject the signaling attempt and send a PathErr message.

If Entropy Label is not used on the signaled LSP and there are no contained LSP, then the Largest LSE Microflow in the Largest Microflow Capacities field MUST be set to the requested bandwidth of the LSP. The optional Largest IP Microflow and Largest L4 Microflow SHOULD be included and MAY be set to configured minimum values.

If Entropy Label is not used on the signaled LSP and LSP that does not have strict packet order constraints contains other LSP, then the LSP Multipath Attributes TLV advertised by the set of contained LSP MUST be used to set the LSP Multipath Attributes TLV Largest Microflow Capacities values for LSP Multipath Attributes TLV. The value of Largest LSE Microflow, Largest IP Microflow, and Largest L4 Microflow in the LSP Multipath Attributes TLV of the containing LSP cannot be less than the maximum effective value of the same parameter for any contained LSP Multipath Attributes TLV.

If Entropy Label is used on the signaled LSP then the Largest LSE Microflow field is set according to the largest microflow that can result from computing the Entropy Label. This value is the greatest of a set of sources of entropy. A configured value MAY be used for IP, or it MAY be assumed that the largest interface carrying IP could carry a single microflow. For contained LSP which have the Entropy Label Used bit clear, the value in the Largest IP Microflow can be

used if the IP Multipath Allowed bit is set for that LSP and the LSR can make use of the IP information in the label stack. For contained LSP which have the Entropy Label Used bit set, the Largest LSE Microflow value already reflects any prior hashing of IP fields.

If the Entropy Label Required bit is set and the LSP being set up is using Entropy Label, then the Largest IP Microflow and Largest L4 Microflow SHOULD be omitted. All midpoint LSR SHOULD not look for entropy beyond the Entropy Label.

If the LSP being set up is not using Entropy Label, then the Largest IP Microflow and Largest L4 Microflow SHOULD be included unless the Entropy Label Used bit is set for every contained LSP. The Largest IP Microflow and Largest L4 Microflow SHOULD be omitted if hashing on the IP addresses or IP addresses and ports would yield no greater entropy than hashing on the label stack alone.

3.3. Path Computation Constraints

The RSVP-TE extensions provides a set of requirements to be met by the links which the LSP is to traverse. This set of requirements also serves as the basis for path computation constraints and for admission control constraints.

3.3.1. Link Multipath Capabilities and Path Computation

Three cases are considered. An LSP may have strict ordering constraints. An MPLS-TP LSP is an example of an LSP with strict ordering constraints. This first type of LSP is covered in Section 3.3.1.1. An LSP may have no ordering constraints at all other than the constraint that microflows cannot be reordered. This second case is covered in Section 3.3.1.2. The remaining case is where an LSP has no ordering constraints but carries traffic for other LSP which do have ordering constraints. This third case is covered in Section 3.3.1.3.

3.3.1.1. Path Computation with Ordering Constraints

For an MPLS-TP LSP or other LSP with a strict packet ordering constraint, any link or FA for which the Ordered Aggregate Enabled bit and Entropy Label Multipath are both clear MUST be excluded from the path computation. If the Default to Multipath bit is set on a link, then setting the OA Depth field to one will override that default.

Link with the Ordered Aggregate Enabled bit clear and the Entropy Label Multipath bit set MAY be included in the path computation if the LSR is capable of encapsulating an LSP with a strict packet

ordering constraint with a fixed Entropy Label. If the LSR is not capable of adding an ELI and EL, then these links MUST be excluded from the path computation.

3.3.1.2. Path Computation with No Ordering Constraint

For an MPLS LSP which has no constraint on packet ordering except that microflows must remain in order and does not contain other LSP with ordering constraints, any link for which the Multipath Enabled bit is set can be used. If a link is advertised as a Link Bundle and the Multipath Enabled bit is set for the link, the available bandwidth SHOULD be taken from the "Unreserved Bandwidth" rather than the "Maximum LSP Bandwidth" (see [RFC4201]).

For most LSP, the bandwidth requirement of the largest microflow is not known but an upper bound is known. For example if the LSP aggregates pseudowires or other LSP of no more than some maximum capacity or LSP which have signaled a microflow upper bound, then an upper bound on the largest microflow is known. If this upper bound exceeds the "Maximum LSP Bandwidth" of a given link, then that link MUST be excluded from the path computation.

3.3.1.3. Path Computation for MPLS containing MPLS-TP

To carry LSP which have strict packet ordering requirements and do not have the Entropy Label Used flag set as a client within a server LSP that do not have strict packet ordering requirements, Entropy Label must be added at the server layer LSP to traverse any link or FA that has the Multipath Enabled bit set. For these LSP links which have the Multipath Enabled bit set and the Entropy Label Multipath bit clear MUST be excluded from the path computation.

If the LSR is not capable of adding an Entropy Label, then to carry any client LSP with the Entropy Label Used clear and the OA Depth set to a non-zero value, the server LSP SHOULD exclude any link or FA that has the Multipath Enabled bit set. For these LSP, any link or FA that has the Multipath Enabled bit set and cannot carry a microflow as large as the entire LSP MUST be excluded from the path computation. These LSP MAY be signaled as having strict packet ordering requirements if they can be carried as a single microflow, but this practice is NOT RECOMMENDED.

3.3.2. Link IP Capabilities and Path Computation

An MPLS-TP LSP cannot be reordered. There may be other types of LSP with strict packet ordering requirements. If LSP with strict packet ordering requirements carry IP, using IP headers in the multipath load distribution would violate the packet ordering requirements.

Some LSP cannot be reordered but do not carry IP, and do not carry payloads which could be mistaken as IP. For example, any LSP carrying only pseudowire traffic, where all pseudowires are using a control word carries no payloads which could be mistaken as IP. These type of LSP can be carried within MPLS LSP that allow use of IP header information in multipath load distribution.

This section focuses on Cases in which links or FA must be excluded from path computation based on the settings of the IP related Flags bits in the Multipath Link Capability TLV.

3.3.2.1. LSP without Packet Ordering Requirements

Many LSP carry only IP or predominantly IP, use no hierarchy or have little diversity in the MPLS label stack, and carry far more traffic than can be carried over a single component link in a multipath. Many LSP due to their high capacity, must traverse only multipath which will use IP header information in the multipath traffic distribution.

For these LSP, links must be excluded from the path computation which do not have the IPv4 Enabled Multipath and IPv6 Enabled Multipath bit set (if carrying both IPv4 and IPv6) and do not have either the Default to IP/MPLS Multipath bit set or the IP Optional Multipath bit set.

Hierarchical PSC LSP which require the use IP header information in the multipath traffic distribution MUST NOT set the Ordered Aggregate Enabled bit, MUST set the Default to IP/MPLS Multipath bit, and MUST NOT set the IP Optional Multipath bit in the FA advertisement. The IP Optional Multipath bit MUST be clear because the LSP cannot change the behavior of midpoint LSR, except perhaps in the case of single hop LSP.

3.3.2.2. LSP with Ordering Requirements

The only time that links or FA with both the Ordered Aggregate Enabled bit and the Entropy Label Multipath bit clear can be used is a special case for MPLS-TP LSP that carry only IP traffic. This case does not apply if the MPLS_TP LSP is carrying other LSP or if it is carrying pseudowires.

Where MPLS-TP LSP are carrying only IP, any link or FA with both the Ordered Aggregate Enabled bit and the Entropy Label Multipath bit clear for which the use of IP header information is not disabled or cannot be disabled on a per LSP basis, that link MUST be excluded from the path computation.

Where MPLS-TP LSP are carrying only IP, links MAY be included in the path computation have the IPv4 Enabled Multipath and IPv6 Enabled Multipath bits clear, or have the Default to IP/MPLS Multipath bit clear, or have the IP Optional Multipath bit set. Links with the IP Optional Multipath set, MUST disable use of IP in the load balance for LSP with the IP Multipath Allowed bit clear.

An MPLS-TP LSP are carrying only IP MUST have OA Depth set to one and LSP Min Depth set to one and the IP Multipath Allowed bit cleared. Call admission SHOULD NOT reject an LSP on the basis of OA Depth being set to one if use of IP headers is always disabled, or can be disabled for the specific LSP. If an MPLS-TP is set up this way and then does start to carry other LSP or carry pseudowires, then reordering within the MPLS-TP LSP will occur.

3.3.3. Link Depth Limitations and Path Computation

For any LSP which does not have strict packet ordering constraints, LSP configuration SHOULD include the following parameters.

LSP Min Depth

a minimal acceptable number of label used in multipath traffic distribution,

LSP IP Depth

a minimal label stack depth where the IP header can be used in multipath traffic distribution

For example, if a PSC LSP will carry LSP which in turn carry very high capacity pseudowires using the pseudowire flow label (see [RFC6391]), the flow label is four labels deep. In this case, LSP Min Depth should be four or higher.

For example, if the same PSC LSP will carry LSP which carry IP traffic with no additional labels, then the IP header is two labels deep. In this case, LSP IP Depth should be two or higher.

For an LSP with non-zero LSP Min Depth, all links with Max Depth set to a value below LSP Min Depth MUST be excluded from the LSP Path Computation.

For an LSP with non-zero LSP IP Depth, all links with IP Depth set to a value below LSP IP Depth MUST be excluded from the LSP Path Computation.

If all LSP carried an accurate LSP Min Depth and LSP IP Depth then neither of these parameters would ever have to be configured. In a network with legacy LSR, it may be necessary to configure these

parameters for some LSP. A per-LSP minimum value of each parameter SHOULD be configurable.

4. Backwards Compatibility

Networks today use three forms of multipath.

1. IP ECMP, including IP ECMP at LER using more than one LSP.
2. Ethernet Link Aggregation [IEEE-802.1AX].
3. MPLS Link Bundling [RFC4201].

4.1. Legacy Multipath Behavior

IP ECMP and Ethernet Link Aggregation always distribute traffic over the entire multipath either using information in the MPLS label stack, or using information in a potential IP header, or using both types of information.

One of two behaviors is assumed for link bundles. Either the link bundles place each LSP in its entirety on a single link bundle component link for all LSP, or link bundles distribute traffic over the entire link bundle using the same techniques used for ECMP and Ethernet Link Aggregation. This second behavior is known as the "all ones" component link (see [RFC4201]).

4.2. Networks without Multipath Extensions

Networks exist that are comprised entirely of LSR which do not support these multipath extensions. In these networks there is no way of telling how multipath links will behave. Since an Ethernet Link Aggregation Group (LAG) is advertised as an ordinary link, there is no way to tell that it is a LAG and not an ordinary link.

4.2.1. Networks with MP Capability on all Multipath

Most large core network today rely heavily on the use of multipath. Ethernet Link Aggregation is heavily used and LSR are configured to use the "all ones" component link for all LSP. The "all ones" component link is the default for many Link Bundling implementations used in core networks.

This is equivalent to the following setting in the Multipath Node Capabilities sub-TLV or Multipath Link Capabilities sub-TLV.

1. Clear the Ordered Aggregate Enabled bit and the IP Optional Multipath bit.
2. Set the Multipath Enabled bit, set the Default to Multipath bit, and clear the Entropy Label Multipath bit.
3. If the label stack is used in the multipath traffic distribution, set Max Depth to the number of label stack entries supported, otherwise set it to zero.
4. Since Entropy Label support is not yet widespread, most LSR would behave as if Entropy Label Multipath were clear.
5. If an IP packet under the label stack can be used in the multipath traffic distribution (very common, almost universal in core LSR), set IPv4 Enabled Multipath, set IPv6 Enabled Multipath, set Default to IP/MPLS Multipath, and set IP Depth to the maximum number of label stack entries which can be skipped over before finding the IP stack. Otherwise clear IPv4 Enabled Multipath, clear IPv6 Enabled Multipath and clear Default to IP/MPLS Multipath.
6. On core networks where UDP and TCP ports are rarely used in multipath, clear all UDP and TCP related bits. On networks where multipath is configured to use TCP and UDP port numbers, these bits would be set.

These networks can support very large LSP but cannot support LSP which require strict packet ordering with other labels below such an LSP, such as pseudowire labels. They may also misroute OAM packet which use GAL (see [RFC5586]) if they use the GAL label in determining the load balance. Generally the Link Bundle advertisements indicate a "Maximum LSP Bandwidth" that is equal to the "Unreserved Bandwidth" if Link Bundling is used with the all-ones component link.

Good or bad, if the behavior is consistent, defaults can be configured in other LSR, such that an accurate guess can be made when no Multipath Link Capability TLV is available for a given link.

For example, in many networks, any link of 10 Gb/s or less can be assumed to be a plain link (no multipath) while any link with a capacity greater than 10 Gb/s can be assumed to be a multipath. These assumptions would hold if no 40 Gb/s or 100 Gb/s links are used.

4.2.2. Netowrks with OA Capability on all Multipath

Some networks, particularly edge networks which tend to be lower capacity, do not use Link Aggregation, and if they use Link Bundling at all, configure each LSR to place each LSP in its entirety on a single link bundle component link for all LSP. Some edge equipment only supports this link bundle behavior.

This is equivalent to the following setting in the Multipath Node Capabilities sub-TLV or Multipath Link Capabilities sub-TLV.

Set the Ordered Aggregate Enabled bit,

Clear the Multipath Enabled bit.

All remaining bits in the Flags field should be clear.

The Max Depth and IP Depth should be set to zero.

These networks can support LSP which require strict packet ordering, but cannot support very large LSP.

Like the behavior described in Section 4.2.1, whether this behavior is good or bad, defaults can be configured which accurately guess the capabilities of links for which no Multipath Link Capability TLV is available.

4.2.3. Legacy Netowrks with Mixed MP and OA Links

Some network may support Ethernet Link Aggregation and all or a subset of LSR which place each LSP in its entirety on a single link bundle component link for all LSP.

If the "Maximum LSP Bandwidth" is set as described in Section 4.2.1, then very large LSP can be supported over link bundles. Very large LSP cannot be supported on LSR which place each LSP in its entirety on a single link bundle component link for all LSP, but these are clearly indicated in signaling,

In these mixed networks it may not be possible to reliably support LSP which require strict packet ordering. It is not possible to know where Ethernet Link Aggregation is used and it is not possible to accurately determine Link Bundling behavior on link bundles where "Maximum LSP Bandwidth" is equal to "Unreserved Bandwidth".

Upgrading LSR to support Entropy Label on both LAG and Link Bundles would improve the ability to carry LSP which require strict packet ordering. To gain any benefit the LSP ingress would have to add ELI

and EL.

If not all LSR are upgraded, then the MPLS TE multipath extensions identify those LSR and multipath that have been upgraded.

4.3. Transition to Multipath Extension Support

If a Multipath Node Capability sub-TLV is not advertised (see Section 2.1), then the LSR does not support these multipath extensions. This allows detection of such nodes and if necessary application of defaults to cover legacy multipath such as typical Ethernet Link Aggregation Behavior.

4.3.1. Simple Transitions

For networks with LSR that do not support multipath extensions, transition is easiest if all legacy LSR support and are configured with a common link bundling behavior. If Ethernet Link Aggregation is not used, a single configured default is needed to cover LSR that do not advertise a Multipath Node Capability sub-TLV.

If Ethernet Link Aggregation had been previously used on Legacy LSR, if possible LAG should be disabled and the members of the former LAG configured and advertised as a link bundle which uses the equivalent "all ones" behavior.

If Ethernet Link Aggregation remains but can be identified in some way, such as links with capacity in excess of some value (for example: greater than 10 Gb/s), then defaults can be set up for these LAG. Alternately administrative attributes could be used [RFC3209].

The transition network in this case lacks the ability to determine the largest microflow that can pass through legacy nodes, but this was the case prior to transition for the entire network prior to transition.

4.3.2. More Challenging Transitions

Transition is made more difficult if legacy LSR in a network support Ethernet Link Aggregation but do not support Link Bundle and cannot be identified by simple means, or the newer LSR lack sufficient configuration capability to support conditional defaults.

This situation is most easily handled if a small upgrade to such an LSR can advertise a fixed Multipath Node Capability sub-TLV giving the characteristics of the Ethernet Link Aggregation implementation on that node. Absent of such cooperation, the problem can be solved by configuration on newer LSR which allows association of a Multipath

Node Capability sub-TLV with a specific legacy router ID and possibly a legacy router ID and link.

LSR supporting Multipath Extensions will need to assign default values through configuration to these legacy LSR running Ethernet Link Aggregation. These default values serve to allow LSP which require strict packet ordering to avoid these legacy LSR.

LSR which do not support [RFC4201] may be sufficiently rare that the ability to assign default values per legacy LSR or per [RFC3209] administrative attribute may not be needed in practice.

5. IANA Considerations

[... to be completed ...]

The symbolic constants IANA-TBD-1 through IANA-TBD-3 need to be replaced. Complete instructions, including identification of the number space for each of these will be added to a later version of this internet-draft.

6. Security Considerations

The combination of MPLS, MPLS-TP, and multipath does not introduce any new security threats. The security considerations for MPLS/GMPLS and for MPLS-TP are documented in [RFC5920] and [I-D.ietf-mpls-tp-security-framework].

7. References

7.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC3471] Berger, L., "Generalized Multi-Protocol Label Switching (GMPLS) Signaling Functional Description", RFC 3471, January 2003.
- [RFC4201] Kompella, K., Rekhter, Y., and L. Berger, "Link Bundling in MPLS Traffic Engineering (TE)", RFC 4201, October 2005.
- [RFC5420] Farrel, A., Papadimitriou, D., Vasseur, JP., and A. Ayyangarps, "Encoding of Attributes for MPLS LSP Establishment Using Resource Reservation Protocol Traffic

Engineering (RSVP-TE)", RFC 5420, February 2009.

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

7.2. Informative References

- [I-D.ietf-mpls-entropy-label]
Kompella, K., Drake, J., Amante, S., Henderickx, W., and L. Yong, "The Use of Entropy Labels in MPLS Forwarding", draft-ietf-mpls-entropy-label-06 (work in progress), September 2012.
- [I-D.ietf-mpls-tp-security-framework]
Fang, L., Niven-Jenkins, B., Mansfield, S., and R. Graveman, "MPLS-TP Security Framework", draft-ietf-mpls-tp-security-framework-04 (work in progress), July 2012.
- [I-D.villamizar-mpls-tp-multipath]
Villamizar, C., "Use of Multipath with MPLS-TP and MPLS", draft-villamizar-mpls-tp-multipath-03 (work in progress), October 2012.
- [IEEE-802.1AX]
IEEE Standards Association, "IEEE Std 802.1AX-2008 IEEE Standard for Local and Metropolitan Area Networks - Link Aggregation", 2006, <<http://standards.ieee.org/getieee802/download/802.1AX-2008.pdf>>.
- [ITU-T.G.800]
ITU-T, "Unified functional architecture of transport networks", 2007, <<http://www.itu.int/rec/T-REC-G/recommendation.asp?parent=T-REC-G.800>>.
- [RFC2475] Blake, S., Black, D., Carlson, M., Davies, E., Wang, Z., and W. Weiss, "An Architecture for Differentiated Services", RFC 2475, December 1998.
- [RFC2702] Awduche, D., Malcolm, J., Agogbua, J., O'Dell, M., and J. McManus, "Requirements for Traffic Engineering Over MPLS", RFC 2702, September 1999.
- [RFC2991] Thaler, D. and C. Hopps, "Multipath Issues in Unicast and Multicast Next-Hop Selection", RFC 2991, November 2000.
- [RFC3209] Awduche, D., Berger, L., Gan, D., Li, T., Srinivasan, V.,

and G. Swallow, "RSVP-TE: Extensions to RSVP for LSP Tunnels", RFC 3209, December 2001.

- [RFC3260] Grossman, D., "New Terminology and Clarifications for Diffserv", RFC 3260, April 2002.
- [RFC3945] Mannie, E., "Generalized Multi-Protocol Label Switching (GMPLS) Architecture", RFC 3945, October 2004.
- [RFC4206] Kompella, K. and Y. Rekhter, "Label Switched Paths (LSP) Hierarchy with Generalized Multi-Protocol Label Switching (GMPLS) Traffic Engineering (TE)", RFC 4206, October 2005.
- [RFC4385] Bryant, S., Swallow, G., Martini, L., and D. McPherson, "Pseudowire Emulation Edge-to-Edge (PWE3) Control Word for Use over an MPLS PSN", RFC 4385, February 2006.
- [RFC5586] Bocci, M., Vigoureux, M., and S. Bryant, "MPLS Generic Associated Channel", RFC 5586, June 2009.
- [RFC5920] Fang, L., "Security Framework for MPLS and GMPLS Networks", RFC 5920, July 2010.
- [RFC6107] Shiimoto, K. and A. Farrel, "Procedures for Dynamically Signaled Hierarchical Label Switched Paths", RFC 6107, February 2011.
- [RFC6391] Bryant, S., Filsfils, C., Drafz, U., Kompella, V., Regan, J., and S. Amante, "Flow-Aware Transport of Pseudowires over an MPLS Packet Switched Network", RFC 6391, November 2011.

Author's Address

Curtis Villamizar (editor)
Outer Cape Cod Network Consulting

Email: curtis@ocnc.com

Network working group
Internet Draft
Category: Standard Track

X. Xu
Huawei
M. Eubanks
AmericaFree.TV
L. Yong
Z. Li
Huawei
N. Sheth
Juniper
Y. Fan
China Telecom

Expires: April 2013

October 8, 2012

Encapsulating MPLS in UDP

draft-xu-mpls-in-udp-03

Status of this Memo

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

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

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

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

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

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

Copyright Notice

Copyright (c) 2009 IETF Trust and the persons identified as the document authors. All rights reserved.

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

Abstract

This document specifies one additional IP-based encapsulation technology for MPLS packets referred to as MPLS-in-UDP, which is intended to facilitate load-balancing the traffic of various MPLS applications such as MPLS-based L2VPN and L3VPN in the core of IP-enabled packet switch networks.

Conventions used in this document

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

Table of Contents

| | |
|---|---|
| 1. Introduction | 3 |
| 2. Terminology | 4 |
| 3. Encapsulation in UDP | 4 |
| 4. Signaling for Encapsulation in UDP | 5 |
| 5. Processing Functions | 5 |
| 6. Applicability | 6 |
| 7. Security Considerations | 6 |
| 8. IANA Considerations | 6 |
| 9. Acknowledgements | 7 |
| 10. References | 7 |
| 10.1. Normative References | 7 |
| 10.2. Informative References | 7 |
| Authors' Addresses | 8 |

1. Introduction

Equal Cost Multi-Path (ECMP) and Link Aggregation Group (LAG) are widely used in the core of IP-enabled Packet Switch Networks (PSN) for load-balancing purposes. Most core routers (i.e., P routers) in the IP-enabled PSN are capable of load-balancing IP traffic flows across ECMP paths and/or LAG based on the hash of the five-tuple of UDP/TCP packets (i.e., source IP address, destination IP address, source port, destination port, and protocol) or some fields in the IP header of non-UDP/TCP packets (e.g., source IP address, destination IP address). However, with existing IP-based encapsulation methods as defined in [RFC4023] (e.g., MPLS-in-IP and MPLS-in-GRE), distinct customer traffic flows of various MPLS applications (e.g., MPLS-based L2VPN or L3VPN) between a given PE pair would be encapsulated with the same IP or GRE tunnel header prior to traversing the IP core. Since the encapsulating traffic is neither TCP nor UDP traffic, core routers could only perform hash calculation on the fields in the IP header of IP or GRE tunnels. As a result, core routers could not achieve an effective load-balancing for these traffic flows in the network due to the lack of adequate entropy information. In most service providers' backbones, MPLS forwarding capability is enabled by default and therefore the deployment of IP-based encapsulation method for MPLS packets (e.g., MPLS-in-IP and MPLS-in-GRE) is not popular. As a result, the above load-balancing issue is unweighted. However, in most cloud data center network environments, data center operators tend to enable IP forwarding capability, rather than MPLS forwarding capability in the underlying data center networks due to certain reasons. In case MPLS-based L2VPN or L3VPN technology are adopted as a scalable data center network solution to support multi-tenancy in such environments, IP-based encapsulation method for MPLS packets would have to be used and therefore the above load-balancing issue would become significant.

[RFC5640] describes a method for improving the load-balancing in Software mesh networks [RFC5565]. However, this method requires core routers to be able to perform hash calculation on the fields including the "load-balancing" field contained in the L2TPv3 or GRE tunnel header. [Entropy-Label] proposes to use the "entropy labels" for achieving a better load-balancing for MPLS traffic flows in the core of MPLS-enabled PSN. Although the entropy label could be inserted in the "Key" field of the GRE header by ingress PE routers in the case where the PSN is IP enabled rather than MPLS enabled, it still requires core routers to be capable of performing hash calculation on the "entropy label" contained in the GRE tunnel

header. Any of the above load-balancing methods requires a change to the data plane of core routers.

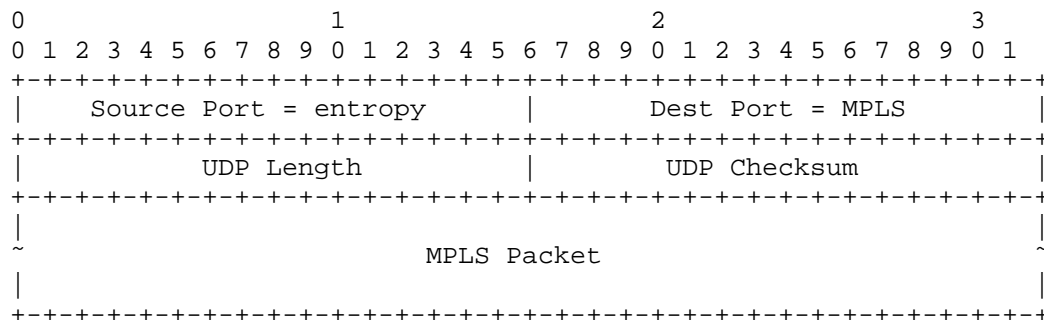
This document describes a new IP-based encapsulation method for MPLS packets referred to as MPLS-in-UDP, which is intended to facilitate load-balancing the traffic of various MPLS applications such as MPLS-based L2VPN and L3VPN in the core of IP-enabled packet switch networks where the core routers could not be upgraded due to some reason.

2. Terminology

This memo makes use of the terms defined in [RFC4364] and [RFC4664].

3. Encapsulation in UDP

MPLS-in-IP messages have the following format:



Source Port of UDP

This field contains an entropy value that is generated by the ingress PE router. For example, the entropy value can be generated by performing hash calculation on certain fields in the customer packets (e.g., the five tuple of UDP/TCP packets). To ensure that the source port number is always in the range 49152 to 65535 which may be required in some cases, instead of calculating a 16-bit hash, the ingress PE router could calculate a 14-bit hash and use those 14 bits as the least significant bits of the source port field while the most significant two bits would be set to binary 11. That still conveys 14 bits of entropy information which would be enough as well in practice.

Destination Port of UDP

This field is set to a value (TBD) indicating the MPLS packet encapsulated in the UDP header is a MPLS unicast one or a MPLS multicast one.

UDP Length

The usage of this field is in accordance with the current UDP specification.

UDP Checksum

The usage of this field is in accordance with the current UDP specification. To simplify the operation on egress PE router, this field is recommended to be set to zero.

4. Signaling for Encapsulation in UDP

PE routers could signal the UDP tunnel encapsulation information among them by some means.

In the case when BGP is used in the MPLS applications (e.g., BGP/MPLS IP VPN [RFC4364]), the MPLS-in-UDP encapsulation information can be signaled by using the mechanism defined in [RFC 5512]. In this case, a new Tunnel Type code for UDP tunnel technology needs to be assigned by IANA. If there is no explicit encapsulation information to signal using the Encapsulation SAFI for the UDP tunneling protocol, a BGP Encapsulation Extended Community with the Tunnel Type set to the value indicating UDP tunneling protocol would be enough. For example, such extended community could be attached to the update messages for NLRI announcement in the BGP/MPLS IP VPN case, or be attached to the update messages dedicated for auto-discovery in the VPLS [RFC4761, RFC4762] case where BGP-based auto-discovery is used. Otherwise, if more detailed information about the UDP tunnel technology is needed for signaling (e.g., to specify what MPLS application is allowed to use this MPLS-in-UDP encapsulation), a new TLV and even a set of sub-TLVs dedicated for UDP tunnel encapsulation technology that would be contained in the Tunnel Encapsulation attribute needs to be defined.

More details about how to signal the MPLS-in-UDP encapsulation information will be described in a separate document.

5. Processing Functions

This MPLS-in-UDP encapsulation causes MPLS packets to be forwarded through "IP UDP tunnels". When performing MPLS-in-UDP encapsulation

by an ingress PE router, the entropy value would be generated by the ingress PE router and then be filled in the Source Port field of the UDP header.

P routers, upon receiving these UDP encapsulated packets, could balance these packets based on the hash of the five-tuple of UDP packets.

Upon receiving these UDP encapsulated packets, egress PE routers would decapsulate them by removing the UDP headers and then process them accordingly.

6. Applicability

Besides the MPLS-based L3VPN [RFC4364] and L2VPN [RFC4761, RFC4762] [E-VPN] applications, MPLS-in-UDP encapsulation could also be used in other MPLS applications including but not limited to 6PE [RFC4798] and PWE3 services.

7. Security Considerations

Just like MPLS-in-GRE and MPLS-in-IP encapsulation formats, the MPLS-in-UDP encapsulation format defined in this document by itself cannot ensure the integrity and privacy of data packets being transported through the MPLS-in-UDP tunnels and cannot enable the tunnel decapsulators to authenticate the tunnel encapsulator. In the case where any of the above security issues is concerned, the MPLS-in-UDP tunnels SHOULD be secured with IPsec in transport mode. In this way, the UDP header would not be seeable to P routers anymore. As a result, the meaning of adopting MPLS-in-UDP encapsulation format as an alternative to MPLS-in-GRE and MPLS-in-IP encapsulation formats is lost. Hence, MPLS-in-UDP encapsulation format SHOULD be used only in the scenarios where all the security issues as mentioned above are not significant concerns. For example, in a data center environment, the whole network including P routers and PE routers are under the control of a single administrative entity and therefore there is no need to worry about the above security issues.

8. IANA Considerations

Two distinct UDP destination port numbers indicating MPLS and MPLS with upstream-assigned label respectively need to be assigned by IANA.

9. Acknowledgements

Thanks to Shane Amante, Dino Farinacci, Keshava A K, Ivan Pepelnjak, Eric Rosen, Kireeti Kompella, Weiguo Hao, Zhenxiao Liu and Xing Tong for their valuable comments on the idea of MPLS-in-UDP encapsulation.

10. References

10.1. Normative References

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

10.2. Informative References

- [RFC4364] Rosen, E and Y. Rekhter, "BGP/MPLS IP Virtual Private Networks (VPNs)", RFC 4364, February 2006.
- [RFC4664] Andersson, L. and Rosen, E. (Editors), "Framework for Layer 2 Virtual Private Networks (L2VPNs)", RFC 4664, Sept 2006.
- [RFC4023] Worster, T., Rekhter, Y., and E. Rosen, "Encapsulating MPLS in IP or GRE", RFC4023, March 2005.
- [RFC5640] Filsfils, C., Mohapatra, P., and C. Pignataro, "Load-Balancing for Mesh Softwires", RFC 5640, August 2009.
- [RFC6391] Bryant, S., Filsfils, C., Drafi, U., Kompella, V., Regan, J., and S. Amante, "Flow Aware Transport of Pseudowires over an MPLS Packet Switched Network", RFC6391, November 2011
- [Entropy-Label] Kompella, K., Drake, J., Amante, S., Henderickx, W., and L. Yong, "The Use of Entropy Labels in MPLS Forwarding", draft-ietf-mpls-entropy-label-01, work in progress, October, 2011.
- [RFC5512] Mohapatra, P. and E. Rosen, "The BGP Encapsulation Subsequent Address Family Identifier (SAFI) and the BGP Tunnel Encapsulation Attribute", RFC 5512, April 2009.
- [RFC4798] J Declercq et al., "Connecting IPv6 Islands over IPv4 MPLS using IPv6 Provider Edge Routers (6PE)", RFC4798, February 2007.

[RFC4761] Kompella, K. and Y. Rekhter, "Virtual Private LAN Service (VPLS) Using BGP for Auto-Discovery and Signaling", RFC 4761, January 2007.

[RFC4762] Lasserre, M. and V. Kompella, "Virtual Private LAN Service (VPLS) Using Label Distribution Protocol (LDP) Signaling", RFC 4762, January 2007.

[E-VPN] Aggarwal et al., "BGP MPLS Based Ethernet VPN", draft-ietf-12vpn-evpn-00.txt, work in progress, February, 2012.

Authors' Addresses

Xiaohu Xu
Huawei Technologies,
Beijing, China

Phone: +86-10-60610041
Email: xuxiaohu@huawei.com

Marshall Eubanks
AmericaFree.TV LLC
P.O. Box 141
Clifton, Virginia 20124
USA

Phone: +1-703-501-4376
Email: marshall.eubanks@gmail.com

Lucy Yong
Huawei USA
1700 Alma Dr. Suite 500
Plano, TX 75075
US

Email: lucyyong@huawei.com

Nischal Sheth
Juniper Networks
1194 North Mathilda Avenue
Sunnyvale, CA 94089 USA

Email: nsheth@juniper.net

Zhenbin Li
Huawei Technologies,
Beijing, China

Phone: +86-10-60613676
Email: lizhenbin@huawei.com

Yongbing Fan
China Telecom
Guangzhou, China.

Phone: +86 20 38639121
Email: fanyb@gsta.com

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: April 25, 2013

Quintin Zhao
Tao Chou
Huawei Technology
Boris Zhang
Telus Communications
Emily Chen
October 22, 2012

P2MP Based mLDP Node Protection Mechanisms for Label Distribution
Protocol P2MP/MP2MP Label Switched Paths
draft-zhao-mpls-mldp-protections-03.txt

Abstract

Existing techniques provide a Point-to-point (P2P) Label Switch Path (LSP) protection mechanism for mLDP nodes. In situations where the data duplication along the p2p backup path is not acceptable, a Point-To-Multipoint (P2MP) or Multipoint-To-Multipoint (MP2MP) LSPs is needed, instead of a P2P LSPs, for the protection of mLDP nodes.

This document defines procedures and protocol extensions for protection of mLDP nodes within Multi-Protocol Label Switching (MPLS) networks using P2MP and MP2MP backup LSPs.

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, 2013.

Copyright Notice

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

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

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

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

Table of Contents

| | |
|---|----|
| 1. Terminology | 4 |
| 2. Requirement Language | 4 |
| 3. Introduction | 5 |
| 3.1. Requirements | 6 |
| 3.2. Scope | 7 |
| 4. mLDP Node Protection using mLDP LSPs | 7 |
| 4.1. Signaling Procedures for P2MP Based Node Protection | 8 |
| 4.1.1. Example of P2MP Based Node Protection's Procedure | 8 |
| 4.1.2. Choose backup upstream LSR | 10 |
| 4.1.3. Create backup path by MRT | 11 |
| 4.1.4. PLR Switching Over Considerations | 12 |
| 4.1.5. mLDP End-to-End Protection | 12 |
| 4.2. Signaling Procedures for MP2MP Based Node Protection | 13 |
| 4.3. Protocol Extensions for mLDP Based Node Protection | 15 |
| 4.3.1. mLDP Based MP Protection Capability Parameter TLV | 15 |
| 4.3.2. mLDP Based MP Node Protection Status Elements | 16 |
| 4.3.3. mLDP Backup FEC Element Encoding | 16 |
| 5. Signaling Procedures for mLDP Based Facility Node Protection | 18 |
| 6. IANA Considerations | 19 |
| 7. Manageability Considerations | 19 |
| 8. Security Considerations | 19 |
| 9. Acknowledgements | 19 |
| 10. References | 20 |
| 10.1. Normative References | 20 |
| 10.2. Informative References | 20 |
| Authors' Addresses | 21 |

1. Terminology

This document uses terminology discussed in [RFC6388] and [I-D.ietf-mppls-ldp-multi-topology]. Additionally the following section provides further explanation for key terms and terminology:

- o PLR: The node where the traffic is logically redirected onto the preset backup path is called Point of Local Repair (PLR).
- o MP: The node where the backup path merges with the primary path is called Merge Point (MP).
- o N: The node to be protected.
- o Pn: The nodes on the backup path for protecting node N.
- o MT-ID: A 16 bit value used to represent the Multi-Topology ID.
- o Default MT Topology: A topology that is built using the MT-ID default value of 0.
- o MT Topology: A topology that is built using the corresponding MT-ID.
- o cut-link: A link whose removal partitions the network. A cut-link by definition must be connected between two cut-vertices. If there are multiple parallel links, then they are referred to as cut-links in this document if removing the set of parallel links would partition the network.
- o cut-vertex: A vertex whose removal partitions the network.
- o MRT: Maximally Redundant Trees. A pair of trees where the path from any node X to the root R along the first tree and the path from the same node X to the root along the second tree share the minimum number of nodes and the minimum number of links. Each such shared node is a cut-vertex. Any shared links are cut-links. Any RT is an MRT but many MRTs are not RTs.

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's demands, operators and service providers continue to deploy multicast applications using Multicast LDP (mLDP) across MPLS networks. For the real-time applications, such as stock trading, on-line games, and multimedia teleconferencing, traditional IGP-mLDP convergence mechanisms fail to meet protection switching times required to minimize, or negate entirely, application interruptions.

Current best practice for protecting services, and subsequently higher-layer applications, include the pre-computation and establishment of a backup path. Once a failure has been detected on the primary path, the traffic will 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, resource consumption and network efficiency.

For a primary LDP P2MP/MP2MP LSP, there are several methods to set up a backup path, these include:

- o The use of an RSVP-TE P2P tunnel as a logical out-going interface, consequently utilize the mature high availability technologies of RSVP-TE.
- o The use of an LDP P2P LSP as a packet encapsulation, so that the complex configuration of P2P RSVP-TE can be skipped.
- o Creating a P2MP/MP2MP backup LSP according to IGP's loop-free alternative route. This solution avoids unnecessary packet duplication compare to the use of a P2P LSP (which is specified in the draft of I-D.wijnands-mpls-mldp-node-protection) and can have 100% scenario coverage if using with multi topology technology, where the backup topology either can be statically configured or dynamically computed/signaled mechanisms such the the mechanism specified in the draft of [I-D.ietf-rtgwg-mrt-frr-architecture].
- o Creation of Multiple Topology (MT) LSP using an entirely disjointed topology.

When the backup path is present, there are two options for packet forwarding and protection switchover:

- o Option 1
The traffic sender transmits the stream on both the primary and

backup path. Once the local traffic receiver detects a failure the switchover will be relatively fast. However the disadvantage of this method is that it consumes bandwidth as duplicate traffic will be sent on the protection and backup path.

- o Option 2

The traffic sender transmits only on the primary path. Although bandwidth resource usage is minimized, cooperation is required to provide adequate switching times and minimize high-layer application impact. Noted that, some mechanisms may need create more than one backup path, like the MRT, and need feed traffics on all the backup paths. That means the MPs need to choose and accept only one traffic of all in such case.

Ideally if switching time performance can be equal or better than the Option 1, it is reasonable to choose option 2 to avoid bandwidth wastage. Some recommendations of this document are based on this viewpoint.

This document specifies P2MP/MP2MP LSP based mLDP node protection.

Note that the computation and configuration of the primary topology and backup topology is out of the scope of this draft, the algorithm can be either MRT based or any other algorithms/method available including the static and offline tools. Besides, how to detect failure is also outside the scope of this document, the mechanism can be bidirectional or unidirectional forwarding detection for link or target object.

3.1. Requirements

A number of requirements have been identified that allow the optimal set of mechanisms to developed. These currently include:

- o Computation of a disjointed (link and node) backup path within the multicast tree;
- o Minimization of protection convergence time;
- o Minimization of operation and maintenance cost;
- o Optimization of bandwidth usage;
- o More protect scenarios can be covered.

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. mLDP Node Protection using mLDP LSPs

By using IGP-FRR or Multi Topology Routing(including the MRT MT routing), LDP can build the backup mLDP LSP among PLR, Pn, and MPs (the downstream nodes of the protected node). In the cases where the amount of downstream nodes are huge, this mechanism can avoid unnecessary packet duplication on PLR, so that can protect the network from traffic congestion risk.

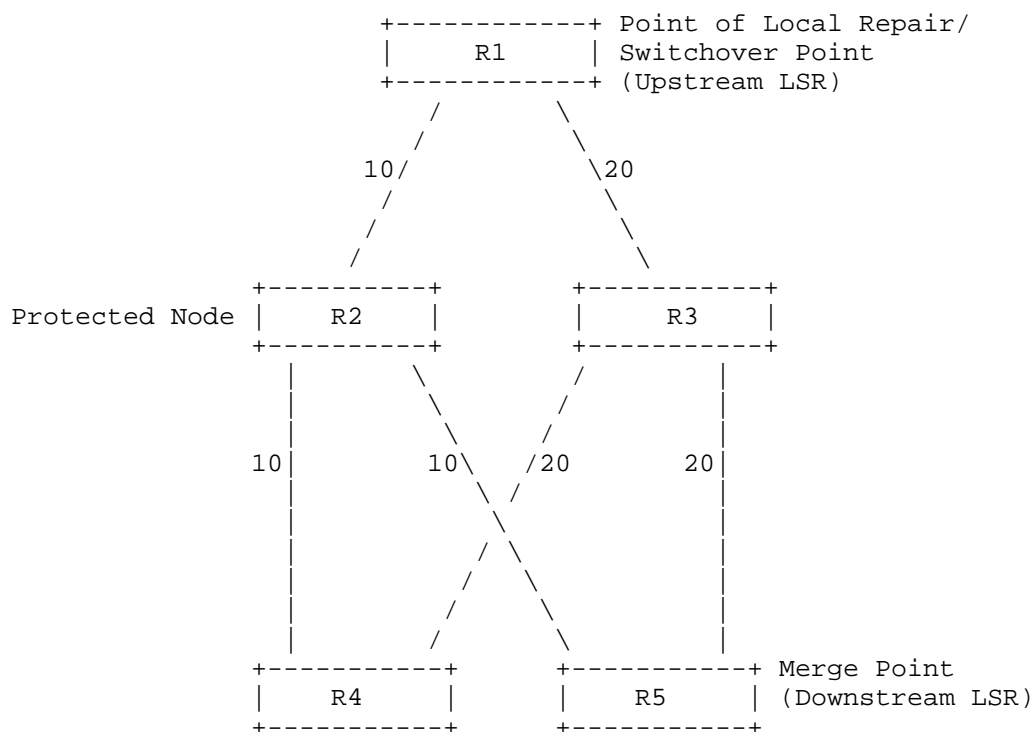


Figure 1: mLDP Local Protection using mLDP LSP Example

In Figure 1, R2 is on the preferential path from R4/5 to R1, 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. This section will take R2 as Protected Node for example, actually the Protected Node can be any Transit node of the mLDP LSP. (We assume that all the nodes in Figure 1 support this mLDP based node protection method, including Pn.)

The procedure of P2MP/MP2MP Based mLDP Node Protection is as follows:

- o As Protected Node, R2 should announce its selected upstream node R1 to all its downstream nodes, which are R4 and R5 in this example. How to know if one node should be protected can be decided by local configuration or its role(transit) in the mLDP LSP.
- o R4 and R5 can consider R1 as the root node of the backup mLDP LSP, and trigger the backup LSP signaling. In parallel, R4/R5 will bind the primary NHLFE(s) to both the backup and primary ILM entry, so that the traffic receiving from backup mLDP LSP can be merged locally to the primary LSP.
- o PLR can distinguish primary LSP and backup LSP by the signaling procedure and only feed traffic on the primary path before failure. When R2 node fails, R1 should switch the traffic to the preset backup path quickly.

In this scenario, if R2 is protected by two P2P LSPs as R1--R3--R4 and R1--R3--R5, the traffic will be duplicated on R1, and R3 will receive two streams. But, If R2 is protected by a mLDP LSP instead, R3 will only receive one stream, and the packet duplication will be done on R3.

The backup mLDP LSP can be P2MP/MP2MP LSP. The P2MP backup LSP is used for P2MP LSP's node protection and the MP2MP backup LSP is used for MP2MP LSP's node protection.

4.1. Signaling Procedures for P2MP Based Node Protection

This section introduces the signaling procedures of P2MP LSP's node protection by backup P2MP LSP.

4.1.1. Example of P2MP Based Node Protection's Procedure

[Editors Note - This section introduces the procedures for P2MP Based Node Protection desires the PLR being capable for node failure detection.]

We assume all the involved nodes have advertised their corresponding

protection capabilities. And the following in this section demonstrates the signaling procedures of P2MP Based Node Protection.

STEP1 Normal procedure of setting up primary path:
Each non-Ingress LSR determine its own upstream LSR and sends out label mapping message, following the procedures as documented in [RFC6388] without any extension. And its upstream LSR will propagate a new label mapping message to its upstream LSR. In such case, we can say the non-Ingress LSR is MP(as R4, R5 in Figure 1), MP's upstream LSR is protected node(as R2 in Figure 1) and protected node's upstream node is PLR(as R1 in Figure 1).

STEP2 Protected Node's procedure of setting up backup path:
After the Protected Node (R2) determines its upstream LSR (R1), it will send the information(PLR's indentify, mLDP FEC) in Notification messages to all its downstream nodes(MPs) immediately. If there are other LSR(s) becoming its downstream node(s) later, it will send such announcement for its new MP(s).

STEP3 MP's procedure of setting up backup path:
When one MP (R4/R5) receive the Notification, it individually determine its secondary paths toward R1 according to the IGP results. Choosing which IGP mechanism's, LFA or MRT etc, results is a local determination. After choosing the backup upstream LSR, MP will send out a FRR Label mapping messages including mLDP backup tree's key <PLR, protected-node, original-mLDP-FEC> and MT-ID if backup path is not in the default topology. Noted that, the label assigned for primary path and secondary path MUST be different to avoid the MP feeding the primary traffic to its secondary path's downstream LSRs. In addition, the original-mLDP-FEC of the backup tree key is encoded in a special opaque value as introduced in section 4.2.3.

STEP4 Pn's procedure of setting up backup path:
When one node receives such aforementioned FRR label mapping message, if it is not the PLR, it can consider itself as a Pn node and will choose its backup upstream node toward PLR on the corresponding topology's shortest IGP path. To avoid the backup LSP going through the Protected Node, additional path selection rule(s) should be applied. A simple method is that the transit nodes can not choose the specified Protected Node as its upstream LSR for the backup LSP. Other methods, such as not-via policy, are under study, and will be added in the future. In order to make the primary and backup topologies rooted from PLR to satisfy the 'maximum disjointed'

requirement, they can be either configured through static configurations or be signaled dynamically through other mechanisms such as MRT.

- STEP5 PLR's procedures of setting up backup path:
When PLR(R1) receives the FRR label mapping message, it can identify that it is the PLR by the mLDP backup FEC elements, so it will decode the special opaque value(which contains the primary mLDP FEC element, introduced in section 4.2.3) and generate the backup forwarding entry for the specific LSP, which is identified by the root address and opaque value in the special opaque value, and bind the backup forwarding state to the specific primary entry, which is indicated by the Protected Node address in the message. Note that there might be more than one backup forwarding entries for a specific protected node.
- STEP6 PLR's procedure when the Protected Node fails:
When failure is detected by PLR, it will switch the traffic to the backup paths. MP will also locally choose to receive which traffic and merge this traffic back to the primary LSP. The switchover manner on PLR is specified in the later section.
- STEP7 Procedure after network re-converges:
When Merge Point(s) see the next hop to Root changed, it/they will advertise the new mapping message(s), and the traffic will re-converge to the new primary path. MP then withdraw the backup label after finishing their re-converge. Pn will delete the specified backup LSP like the procedure of deleting normal P2MP LSP. And the entire backup P2MP LSP will be deleted when all the node MP leave the backup P2MP LSP.

4.1.2. Choose backup upstream LSR

Obviously, the backup path can not go through the protected node N, this section discusses how to choose the backup upstream LSR to avoid N.

Firstly, finding out the candidate upstream LSRs as below:

- o MPs should preferentially choose the upstream LSRs on the shortest path as candidates, except node N. If no other upstream LSRs on the shortest path, MPs should choose the next-hop on N's detour path as candidate. The detour path can be IGP-FRR path or other topologies' disjoint paths. The IGP-FRR path can be provided by LFA, U-Turn, etc. The disjoint path can be provided by MT, MRT, etc. How to choose the candidates is a local decision, can be

determined by configuration.

- o For the Pn node, it MUST choose from the IGP next-hops on the shortest path toward PLR within the topology specified in the FRR mLDP FEC element by MT-ID field. The candidate upstream LSRs MUST except the node N.

Then, each node can choose one from the candidate upstream LSRs as its backup upstream LSR, following the algorithm described in [RFC6388] section 2.4.1.1.

4.1.3. Create backup path by MRT

The algorithm of Maximally Redundant Trees(MRTs), which is defined in [I-D.enyedi-rtgwg-mrt-frr-algorithm], can compute two topologies(Blue and Red) automatically. The two topologies can provide a pair of maximally disjoint paths from one MP to PLR. The failed node will not exist in both of the paths, unless it is the cut-vertex. So these paths can be used as backup paths for the mLDP node protection.

Two backup multicast trees need be created along the MRT paths in each MRT topologies, because sometimes MPs can not indentify which path can avoid the failed node. The tree in one MRT topology also uses the combination of <PLR, failed-node, original-mLDP-FEC> as its own key, and the two trees in different topologies is distinguished by MT-ID. The MRT backup tree's creation uses the same procedure described as above.

The announcement with PLR's information from the protected node triggers the MP to send backup mapping message along the MRT path in both topologies, with corresponding tree's key, labels and MT-ID. One node receives the backup message and find out it is not the PLR, then it send out a new backup mapping message along the corresponding path. If one node's multicast upstream node can only be the protected node, this node can stop the procedure. When PLR receives these messages, it associates the primary tree with the backup tree. PLR may receive two backup trees if both paths can avoid the protected node.

Because one MRT tree may not include all MPs, PLR must feeds the traffics to both corresponding backup trees once PLR detects the failure. And MPs may receive packets from both MRT paths, MPs MUST drop the packets in the Red topology in such case.

MRT makes the solution more complex, but it can be deployed automatically and reach 100 percent scenario coverage in theory.

4.1.4. PLR Switching Over Considerations

The P2MP Based Node Protection also has the BFD scalability issue on the Protected node. Similar with P2P Based Node Protection solution, this section provides two methods for deployment.

- o Option 1:
If PLR cannot differentiate link and node failure, MP must take the responsibility to drop one of the two reduplicate traffics when failure is detected. In this case, the Node Failure Required Flag, in the P2MP Based MP Node Protection Status Element, must be set as 'N'. PLR will switch the traffic to the backup path when failure detected and MP will drop traffic on the backup path until it sees N fails.
- o Option 2:
If PLR can differentiate link and node failure, PLR MUST NOT switch the traffic to the backup path until it detects the node N failure. In this case, the Node Failure Required Flag, in the P2MP Based MP Node Protection Status Element, must be set as 'Y'.

Note that, all the MPs of N MUST use one same Node Failure Required Flag value. Otherwise, the backup P2MP LSP tree need depart to two trees different from the switch over type, and this part is TBD. And it is also possible that can use a backup MP2MP LSP tree to protect one node in the primary MP2MP LSP tree, this part is TBD too

[Editors Note - This Editors note and remaining options will be removed before publication of this document.]

4.1.5. mLDP End-to-End Protection

[I-D.ietf-mpls-ldp-multi-topology] provides the mechanism to setup disjointed LSPs within different topologies. Applications can use these redundant LSPs for end-to-end protection based on MT.

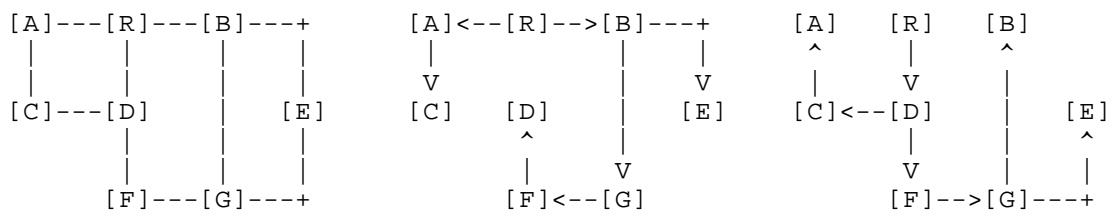
The backup topologies can be build by static configuration or automatic computation. The static method need create the disjoint trees artificially in one topology, root node can setup 1:1 or 1+1 End-to-End protection, using these backup disjoint mLDP LSP. The automatic method can use MRT algorithm. MRT can also provide maximally disjoint P2P paths to build a pair of redundant multicast trees from leaves to root. This section mainly analyses the automatic method.

The procedure of building backup multicast trees by MRT just like the creation of primary multicast tree. Leaf triggers building multicast tree along the path toward root in both MRT topologies, colored as

Blue and Red, separately.

Each MRT backup tree can cover all the leaves, but two different sets of leaves maybe share one same mid node(not the cut-vertex). Therefore, the root must feed traffic to both two MRT trees when failure, and the leaves must drop the packets in the Red topology if receives packets on both MRT backup tree.

Take Figure 2 for example, node R is the root and the other nodes are leaves. When the node G breaks, node F,D will not receive traffic in Blue MRT and node E,B will not receive traffic in Red MRT. Therefore, only feeds traffics in both MRTs can protect all leaves. In addition, node A,C can receive traffics in both MRTs, they must choose dropping the one in Red MRT.



(a) Original topology (b) Blue MRT of root R (c) Red MRT of root R

Figure 2: mLDP End-to-End Protection using MRT

Because MRT computation is separate in different areas, in the case of inter-area, root and leaves are not in the same area, the BR(Border Router) must do some special procedure for protecting another BR. For example, BR1 and BR2 cross the area(x) and area(y). Root is in area(y). One node, LSR1, in area(x) may choose BR1 as its Red MRT upstream LSR to against BR2's failure. But maybe BR1 will choose BR2 as its Red MRT upstream LSR in area(y). LSR1 will receive no traffic when BR2 fails. To solving such problem, BR can choose to treat itself as a leaf when it receives the mLDP mapping message, which need cross areas, in MRT topology. That means BR needs to join both the Blue and Red MRT trees in such case, and drops the Red MRT's traffic if it receives traffic in both MRTs.

In order to reduce the packets' loss in convergence, End-to-End Protection also needs leaves to support MBB procedure.

4.2. Signaling Procedures for MP2MP Based Node Protection

This section introduces the solution to protect MP2MP LSP node by backup MP2MP LSP.

The procedure is similar with the P2MP based node protection. MP send backup label mapping message with MP2MP downstream FRR FEC element. When PLR receives backup label mapping message with MP2MP downstream flag, it send the backup label mapping message with mp2mp upstream FRR FEC element to Pn, and then finally to MPs. This procedure just follows the normal MP2MP LSP procedure.

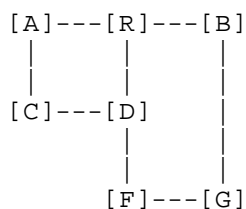
PLR node binds its backup MP2MP downstream NHLFE entry to primary MP2MP downstream ILM entry and binds its primary MP2MP upstream NHLFE entry to backup MP2MP upstream ILM entry.

MP node binds its primary MP2MP downstream NHLFE entry to backup MP2MP downstream ILM entry and binds its backup MP2MP upstream NHLFE entry to primary MP2MP upstream ILM entry.

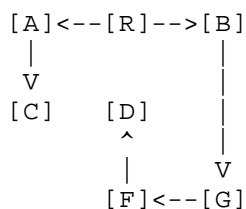
Once detecting the protected node failure, PLR switches the downstream traffic to backup path and MP switches the upstream traffic to backup path.

End-to-End protection can also use the backup multicast tree to protect MP2MP applications. The biggest difference between End-to-End and Node protection is the detecting method, which is outside the scope of this document. In addition, the MRT may not suitable for MP2MP End-to-End protection at present. The disjoint path from leaf to root can not provide protection for the traffic from leaf to leaf. Because the upstream traffic sent from leaf to leaf include two directions, downstream and upstream. One node may be another one's downstream LSR in both directions of both MRT topologies.

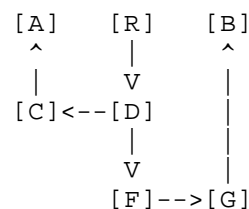
For example in figure 3, node R is the root and the other nodes are leaves of the MP2MP LSPs. The downstream traffic from root to leaves is along the path shown in sub figure (b) and (c), and the upstream traffic from leaf G to other nodes is along the path shown in sub figure (e) and (f). Obviously, if the node F breaks, the node D can not receive the traffic sent by leaf G in both MRT topologies.



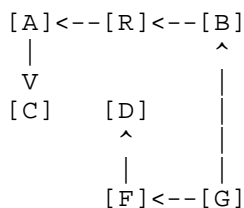
(a) Original topology



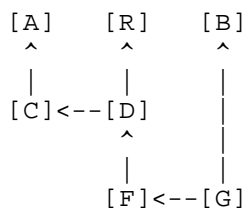
(b) Blue MRT of root R



(c) Red MRT of root R



(e) Blue MRT of root G



(f) Red MRT of root G

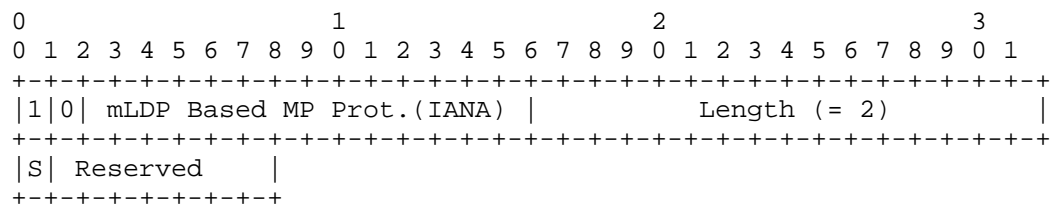
Figure 3: The problem of MP2MP End-to-End Protection using MRT

[TBD]This section is still in research, the details will be shown in the future versions.

4.3. Protocol Extensions for mLDP Based Node Protection

4.3.1. mLDP Based MP Protection Capability Parameter TLV

A new Capability Parameter TLV is defined as mLDP Based MP Protection Capability for node protection. Following is the format of this new Capability Parameter TLV:



S: As specified in [RFC5561]

Figure 4: mLDP Based MP Protection Capability

This is an unidirectional capability announced.

An LSR, which supports the mLDP based protection procedures, should advertise this mLDP Based MP Protection Capability TLV to its LDP speakers. Without receiving this capability announcement, an LSR MUST NOT send any message including the mLDP Based MP Node Protection Status Element and mLDP Backup FEC Element to its peer.

Capability Data might be needed to distinguish the capabilities of different nodes, such as PLR, MP, N, Pn and so on. This part is TBD.

4.3.2. mLDP Based MP Node Protection Status Elements

A new type of LDP MP Status Value Element is introduced, for notifying upstream LSR information. It is encoded as follows:

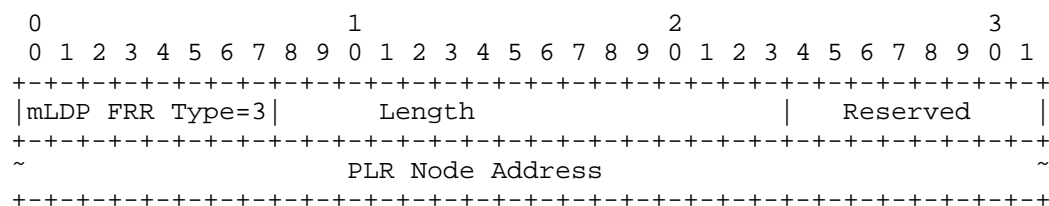


Figure 5: FRR LDP MP Status Value Element

mLDP FRR Type: Type 3 (to be assigned by IANA)

Length: If the Address Family is IPv4, the Length MUST be 5;
if the Address Family is IPv6, the Length MUST be 17.

PLR Node Address: The host address of the PLR Node.

4.3.3. mLDP Backup FEC Element Encoding

A new type of mLDP backup FEC Element is introduced, for notifying upstream LSR information. It is encoded as follows:

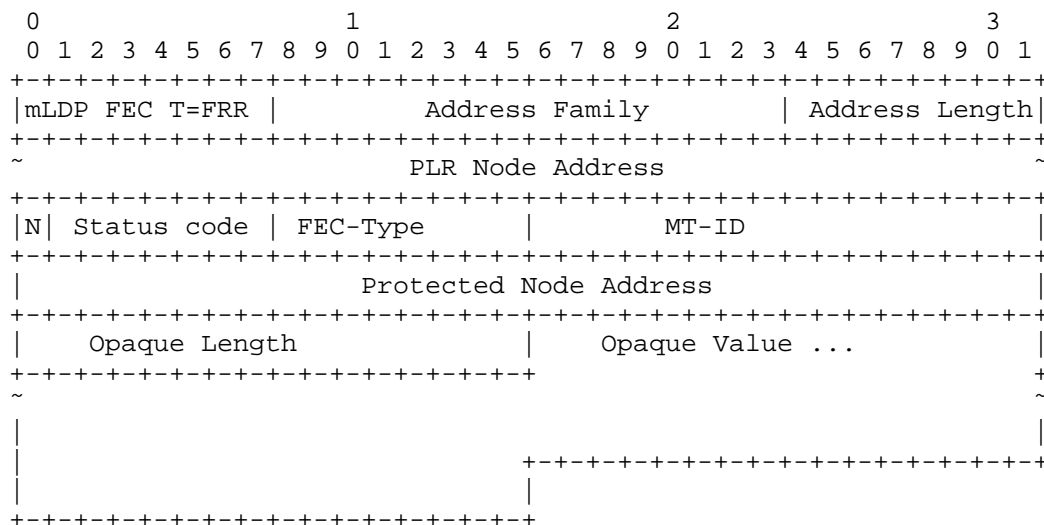


Figure 6: mLDP Backup FEC Element

mLDP FEC Type-FRR: Type 5 (to be assigned by IANA)

Length: If the Address Family is IPv4, the Address Length MUST be 9;
if the Address Family is IPv6, the Address Length MUST be 33.

Status code: 1 = Primary path for traffic forwarding
2 = Secondary path for traffic forwarding

FEC-Type: 6 = P2MP FEC type
7 = MP2MP-up FEC type
8 = MP2MP-down FEC type

PLR Node Address: The host address of the PLR Node.

Protected Node Address: The host address of the Protected Node.

N Bit: Node Failure Required Flag, the occasion of switching traffic's on PLR failure
1 = 'Y', switch traffic to backup path only when PLR detects the node failure
0 = 'N', switch traffic to backup path when PLR detects failure

Opaque Length: The length of the opaque value, in octets.

Opaque Value: One or more MP opaque value elements, the same definition in [RFC6388].

Specially for the FRR mLDP FEC element, the Opaque Value MUST be encoded as the Recursive Opaque Value, which is defined in [RFC6512]. The fields of the Recursive Opaque Value contains the original primary path's mLDP FEC element.

The encoding for this Recursive Opaque Value, as defined in [RFC6512], is shown in Figure 5.

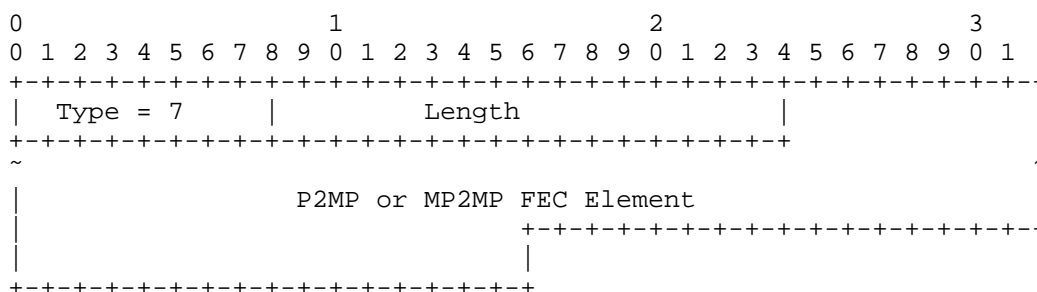


Figure 7: Recursive Opaque Value, defined in [RFC6512]

The Opaque Value is encoded by MP node and decoded by PLR. Other nodes MUST NOT interpret the opaque value at all.

5. Signaling Procedures for mLDP Based Facility Node Protection

[TBD]In the detour solution, one backup LSP protects one primary one. So if there are several mLDP LSPs using one backup path, there will be several backup LSPs on one same backup path. In such case, this may cause one kind waste of LSP resource. Using the facility node protection solution can minimize such waste, the cost is making the procedure more complicated.

MP chooses its primary upstream LSR as N and send label mapping message to N. When N receives the label mapping message from MP, it will assign a upstream label to MP. MP uses this upstream label as its incoming label and release the label resource it used for this LSP before. Then MP will find the backup mLDP LSP by the specified PLR and N address, if no such LSP exists, MP will trigger creating one. The backup mLDP LSP is exclusive by PLR and N address. N uses this upstream label as its own incoming label and send this label's label mapping message to PLR. After PLR create the primary LSP, it will find one backup mLDP LSP by the specified PLR and N address. If there exists one such LSP, PLR will bind the backup LSP to primary

one.

When PLR detects the N's failure, it switches traffic to backup path using dual label stack, the inner label is the outgoing label from N and the outer label is the backup LSP's outgoing label from Pn. MP will receive the traffic from Pn, which is same as the traffic from N.

The key point of the facility solution is node N how to assign the upstream label. This solution is still under research.

6. IANA Considerations

This memo includes the following requests to IANA:

- o mLDP Based MP Protection Capability.
- o mLDP FRR types for LDP MP Status Value Element.
- o mLDP FEC FRR Element type.

7. Manageability Considerations

[Editors Note - This section is TBD.]

8. 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.

Note that authorization policies should be implemented and/or configure at all the nodes involved.

9. Acknowledgements

We would like to thank Nicolai Leymann and Daniel King for his valuable suggestions regarding to this draft. We also would like to

thank Robin Li, Lujun Wan for their comments and suggestions to the 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.
- [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.
- [RFC6388] Wijnands, IJ., Minei, I., Kompella, K., and B. Thomas, "Label Distribution Protocol Extensions for Point-to-Multipoint and Multipoint-to-Multipoint Label Switched Paths", RFC 6388, November 2011.
- [RFC6512] Wijnands, IJ., Rosen, E., Napierala, M., and N. Leymann, "Using Multipoint LDP When the Backbone Has No Route to the Root", RFC 6512, February 2012.

10.2. Informative References

- [I-D.ietf-mpls-ldp-multi-topology]
Zhao, Q., Fang, L., Zhou, C., Li, L., and N. So, "LDP Extensions for Multi Topology Routing",
draft-ietf-mpls-ldp-multi-topology-04 (work in progress),
July 2012.
- [I-D.wijnands-mpls-mldp-node-protection]
Wijnands, I., Rosen, E., Raza, K., Tantsura, J., Atlas, A., and Q. Zhao, "mLDP Node Protection",
draft-wijnands-mpls-mldp-node-protection-01 (work in progress),
June 2012.
- [I-D.ietf-rtgwg-mrt-frr-architecture]

Atlas, A., Kebler, R., Envedi, G., Csaszar, A.,
Konstantynowicz, M., White, R., and M. Shand, "An
Architecture for IP/LDP Fast-Reroute Using Maximally
Redundant Trees", draft-ietf-rtgwg-mrt-frr-architecture-01
(work in progress), March 2012.

[I-D.enyedi-rtgwg-mrt-frr-algorithm]

Atlas, A., Envedi, G., Csaszar, A., and A. Gopalan,
"Algorithms for computing Maximally Redundant Trees for
IP/LDP Fast- Reroute",
draft-enyedi-rtgwg-mrt-frr-algorithm-02 (work in
progress), October 2012.

Authors' Addresses

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

Email: quintin.zhao@huawei.com

Tao Chou
Huawei Technology
156 Beiqing Rd
Haidian District, Beijing 100095
China

Email: tao.chou@huawei.com

Boris Zhang
Telus Communications
200 Consilium Pl Floor 15
Toronto, ON M1H 3J3
Canada

Phone:
Email: Boris.Zhang@telus.com

Emily Chen
2717 Seville Blvd, Apt 1205
Clearwater, FL 33764
US

Email: emily.chen220@gmail.com

